Quantifying stratospheric biases and identifying their potential sources in subseasonal forecast systems

Zachary D. Lawrence[1,2], Marta Abalos[3], Blanca Ayarzagüena[3], David Barriopedro[3], Amy H. Butler[4], Natalia Calvo[3], Alvaro de la Cámara[3], Andrew Charlton-Perez[5], Daniela I.V. Domeisen[6,7], Etienne Dunn-Sigouin[8], Javier García-Serrano[9], Chaim I. Garfinkel[10], Neil P. Hindley[11], Liwei Jia[12,13], Martin Jucker[14], Alexey Y. Karpechko[15], Hera Kim[16], Andrea L. Lang[17], Simon H. Lee[18], Pu Lin[13,19], Marisol Osman[20,21], Froila M. Palmeiro[9], Judith Perlwitz[2], Inna Polichtchouk[22], Jadwiga H. Richter[23], Chen Schwartz[10], Seok-Woo Son[16], Irina Statnaia[15], Masakazu Taguchi[24], Nicholas L. Tyrrell[15], Corwin J. Wright[11], and Rachel W.-Y. Wu[7]

[1]Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, CO, USA
[2]NOAA Physical Sciences Laboratory (PSL), Boulder, CO, USA
[3]Department of Earth Physics and Astrophysics, Universidad Complutense de Madrid, Madrid, Spain
[4]NOAA Chemical Sciences Laboratory (CSL), Boulder, CO, USA
[5]Department of Meteorology, University of Reading, Reading, UK
[6]University of Lausanne, Lausanne, Switzerland
[7]ETH Zurich, Zurich, Switzerland
[8]NORCE Norwegian Research Centre and Bjerknes Centre for Climate Research, Bergen, Norway
[9]Group of Meteorology, Universitat de Barcelona (UB), Barcelona, Spain
[10]Fredy & Nadine Herrmann Institute of Earth Sciences, The Hebrew University of Jerusalem, Israel
[11]Centre for Space, Atmospheric and Oceanic Science, University of Bath, Bath, UK
[12]University Corporation for Atmospheric Research, Boulder, CO
[13]Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, NJ
[14]Climate Change Research Centre and ARC Centre of Excellence for Climate Extremes, University of New South Wales, Sydney, Australia
[15]Finnish Meteorological Institute, Meteorological Research, Helsinki, Finland
[16]School of Earth and Environmental Sciences, Seoul National University, South Korea
[17]Department of Atmospheric and Environmental Sciences, University at Albany, State University of New York, Albany, New York
[18]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY
[19]Program in Atmospheric Science, Princeton University, Princeton, NJ
[20]CONICET – Universidad de Buenos Aires, Centro de Investigaciones del Mar y la Atmósfera (CIMA), Buenos Aires, Argentina
[21]Now at Institute of Meteorology and Climate Research (IMK-TRO), Department Troposphere Research, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[22]European Centre for Medium-Range Weather Forecasts, Reading, UK
[23]Climate and Global Dynamics Laboratory, National Center for Atmospheric Research, Boulder, CO
[24]Department of Earth Science, Aichi University of Education, Kariya, Japan

Correspondence: Zachary D. Lawrence (zachary.lawrence@noaa.gov)

Abstract. The stratosphere can be a source of predictability for surface weather on timescales of several weeks to months. However, the potential predictive skill gained from stratospheric variability can be limited by biases in the representation of

stratospheric processes and the coupling of the stratosphere with surface climate in forecast systems. This study provides a first systematic identification of model biases in the stratosphere across a wide range of subseasonal forecast systems.

It is found that many of the forecast systems considered exhibit warm global mean temperature biases from the lower to middle stratosphere, too strong/cold wintertime polar vortices, and too cold extratropical upper troposphere/lower stratosphere regions. Furthermore, tropical stratospheric anomalies associated with the Quasi-Biennial Oscillation tend to decay toward each system's climatology with lead time. In the Northern Hemisphere (NH), most systems do not capture the seasonal cycle of extreme vortex event probabilities, with an underestimation of sudden stratospheric warming events and an overestimation of strong vortex events in January. In the Southern Hemisphere (SH), springtime interannual variability of the polar vortex is generally underestimated, but the timing of the final breakdown of the polar vortex often happens too early in many of the prediction systems.

These stratospheric biases tend to be considerably worse in systems with lower model lid heights. In both hemispheres, most systems with low-top atmospheric models also consistently underestimate the upward wave driving that affects the strength of the stratospheric polar vortex. We expect that the biases identified here will help guide model development for sub-seasonal to seasonal forecast systems, and further our understanding of the role of the stratosphere for predictive skill in the troposphere.

1 Introduction

The Earth's stratosphere is home to several dynamical phenomena that are coupled with the tropospheric circulation. This coupling between the two layers can go in both directions; tropospheric variability drives variability in the stratosphere, but downward coupling from stratospheric variability can subsequently impact weather in the troposphere across the globe (Domeisen and Butler, 2020). As a result, the stratosphere is recognized as a source of predictability for the troposphere on subseasonal-to-seasonal (S2S) timescales (Butler et al., 2019a; Domeisen et al., 2020b). However, model simulations and forecasts often struggle to adequately capture such stratosphere-troposphere coupling processes. Model biases in both the troposphere and stratosphere can impact these coupling processes, with potential deleterious effects for S2S predictability. The goal of this study is to provide a systematic identification of stratospheric biases in a wide range of S2S forecast systems.

In the wintertime extratropical stratosphere, variability in the upward flux of planetary-scale Rossby waves drives variability in the westerly circulation of the stratospheric polar vortices, affecting their mid-winter strength and the timing of their seasonal breakdowns in spring (Andrews et al., 1987; Garfinkel et al., 2010). During winter and spring, anomalous behavior of the polar vortices can exert an influence on the underlying troposphere. This kind of "downward coupling" is especially apparent for extreme polar vortex events, including sudden stratospheric warming (SSW) events (Gerber et al., 2012; Baldwin et al., 2021), which are characterized by massive disruptions to the polar vortex that decelerate and reverse its westerly winds; strong vortex events, which are characterized by anomalous strengthening of the vortex (Limpasuvan et al., 2005; Tripathi et al., 2015); and final warmings, which denote the breakdown of the polar vortex in spring or early summer until the subsequent autumn (Black et al., 2006). The observed surface response following such polar vortex events generally resembles the so-called Northern and Southern Annular Modes (NAM/SAM; Thompson and Wallace, 2000; Baldwin and Dunkerton, 2001), or the North Atlantic

Oscillation (NAO; Ambaum and Hoskins, 2002; Charlton-Perez et al., 2018; Domeisen, 2019), which captures the tropospheric response specifically over the North Atlantic, where northern hemisphere (NH) stratospheric influence tends to be the strongest (Butler et al., 2017; Dai and Hitchcock, 2021). The surface temperature and precipitation responses associated with these large-scale tropospheric circulation patterns have been shown to contribute to extreme events such as cold air outbreaks and precipitation extremes (Domeisen and Butler, 2020).

In the tropical stratosphere, the absorption of vertically propagating tropical waves from the troposphere below drive the alternating phases of easterly and westerly winds of the Quasi-Biennial Oscillation (QBO; Baldwin et al., 2001). In turn, the QBO has the ability to modulate the tropospheric jet streams, tropical convection, and other phenomena such as the Madden-Julian Oscillation (e.g., Gray et al., 2018; Kim et al., 2020; Anstey et al., 2021, and references therein). These tropospheric teleconnections of the QBO are known to affect the NAO, as well as surface temperatures and precipitation near East Asia and in the tropics on both seasonal-mean and subseasonal timescales (Anstey et al., 2021; Gray et al., 2018; Haynes et al., 2021; Elsbury et al., 2021; Park et al., 2022). The QBO also has an apparent influence on the strength of the polar vortex, particularly in the Northern Hemisphere (NH), whereby easterly or westerly winds in the tropical lower stratosphere tend to lead to a weaker or stronger polar vortex, respectively (e.g., Holton and Tan, 1980; Calvo et al., 2007; Garfinkel et al., 2012; Anstey and Shepherd, 2014; Gray et al., 2018; Rao et al., 2020a). This is known as the "Holton-Tan" effect; there are several mechanisms that help explain such QBO-polar vortex coupling, but generally they are tied to the QBO winds affecting the propagation and dissipation of waves in the polar stratosphere.

Since the stratosphere and its variations generally exhibit a higher persistence and predictability than the troposphere (Domeisen et al., 2020a; Son et al., 2020), their downward influence can lead to opportunities for long-range prediction of surface weather on S2S timescales (Butler et al., 2019a). For example, all of the extreme polar vortex events mentioned above have shown potential for improved S2S surface prediction: SSWs (Sigmond et al., 2013), strong vortex events (Tripathi et al., 2015; Domeisen et al., 2020b), and final warmings (Butler et al., 2019b; Byrne et al., 2019). Tropical stratospheric variability associated with the QBO has also shown the potential to improve surface prediction on these time-scales (Marshall and Scaife, 2009; Garfinkel et al., 2018; Martin et al., 2021b). However, surface prediction skill is not always improved based on the conditions in the stratosphere. In fact, several regions exhibit poorer predictability after extreme polar vortex events as compared to periods without (Domeisen et al., 2020b). There are likely several reasons for these shortcomings in dynamical predictions related to the inherent chaotic nature of dynamical coupling within the atmosphere. For example, not all SSWs appear to have a significant downward impact (Karpechko et al., 2017), likely due to tropospheric internal variability (Domeisen et al., 2020c; Afargan-Gerstman and Domeisen, 2020) and the duration and strength of anomalies in the lower stratosphere following the SSW (Maycock and Hitchcock, 2015; Karpechko et al., 2017; Charlton-Perez et al., 2018; Domeisen, 2019; White et al., 2020).

Model biases and related shortcomings in simulating the stratosphere can affect dynamical predictions of stratospheric variability and stratosphere-troposphere coupling (Domeisen et al., 2020b; Charlton-Perez et al., 2013). Indeed, tropospheric biases identified independently of the stratosphere can sometimes be traced back to stratospheric anomalies in subseasonal forecasts; in the ECMWF extended-range prediction system, the persistence of the NAO was found to be constrained too strongly by the state of the polar vortex at initialization (Kolstad et al., 2020). Since stratospheric variability and extreme events are primarily

governed by wave-mean flow interactions, biases in wave driving due to either resolved or parameterized processes can lead to biases in the mean state of the stratosphere, which can further alter its response to subsequent forcing (McLandress et al., 2012; Richter et al., 2014) and limit stratospheric predictability (Portal et al., 2021). More fundamentally, a model's vertical resolution in the stratosphere can influence its ability to realistically represent the stratosphere. For example, climate models with limited resolution in the stratosphere tend to dramatically underestimate stratospheric variability (Charlton-Perez et al., 2013; Shaw et al., 2014; Richter et al., 2020; Rao et al., 2020a). Similarly, subseasonal forecast systems with severely limited model resolution in the stratosphere exhibit poor skill in predicting extreme polar vortex events beyond 1 week, whereas systems with higher model lids and higher vertical resolution in the stratosphere can predict these events up to 2-3 weeks in advance (Domeisen et al., 2020a). Subseasonal forecast systems that poorly represent tropospheric stationary waves also tend to have low vertical resolution in the stratosphere, connected to larger stratospheric biases (Schwartz et al., 2021). On seasonal timescales, surface prediction skill can also in part be traced back to model properties such as vertical resolution in the stratosphere (Butler et al., 2016; Portal et al., 2021).

It is expected that identifying and properly addressing the sources of stratospheric biases would not only improve prediction in the stratosphere, but also help improve the surface prediction skill associated with the aforementioned stratospheric phenomena. For example, some recent studies have investigated the impact of explicitly correcting stratospheric biases in model simulations, and found improvements to SSW statistics (Tyrrell et al., 2021) and the representation of the Holton-Tan effect (Karpechko et al., 2021). In other similar studies, teleconnections to the polar vortex related to Siberian snow cover (Tyrrell et al., 2020) and the El Niño-Southern Oscillation (ENSO; Tyrrell and Karpechko, 2021) were shown to be sensitive to model stratospheric biases. In the case of the Siberian snow cover, explicitly correcting the stratospheric biases had modest impacts on the magnitude and duration of downward coupling between the polar vortex and the AO/NAO. However, in the case of ENSO teleconnections, correcting stratospheric biases had no detectable impact on the NAO, possibly because the NAO response to the strong ENSO events in the experiments was dominated by the tropospheric ENSO pathway to the North Atlantic (Jiménez-Esteve and Domeisen, 2018).

Given the role the stratosphere can play for skillfully predicting surface weather at subseasonal to seasonal timescales, it is crucial to investigate and diagnose the stratosphere and stratosphere-troposphere coupling biases present in such models. Identifying and comparing these biases will serve as a data point for improving subseasonal forecast systems, and ultimately prediction skill, both in the stratosphere and for extended-range surface weather. Herein we perform a comprehensive investigation and intercomparison of stratosphere-related biases affecting subseasonal prediction systems. The present work is the result of a volunteer collaborative effort of the World Climate Research Programme (WCRP) Stratosphere-troposphere Processes and their Role in Climate (SPARC) Stratospheric Network for the Assessment of Predictability (SNAP) activity, which is also the stratosphere sub-project of the WCRP/World Weather Research Programme S2S Prediction Project. In Section 2 we describe the datasets and methods we use to identify biases. We describe our results in Section 3, and discuss and summarize our findings in Section 4.

2 Data & Methods

2.1 Subseasonal-to-Seasonal (S2S) Hindcast and Reanalysis Datasets

We primarily use ensemble hindcast data from the S2S Prediction Project Database (Vitart et al., 2017). Where possible, we also include results from other ensemble forecast systems that do not provide data to the S2S database; these include the National Oceanic and Atmospheric Administration's Global Ensemble Forecast System version 12 (NOAA GEFSv12; Hamill et al., 2021; Guan et al., 2021), the Geophysical Fluid Dynamics Laboratory Seamless System for Prediction and EArth System Research (GFDL-SPEAR; Delworth et al., 2020), the National Center for Atmospheric Research Community Earth System Model version 2 (CESM2) with version 6 of the Community Atmosphere Model as its atmospheric component (NCAR CESM2-CAM6, hereafter CESM2-CAM or CESM2-C), and CESM2 with the version 6 Whole Atmosphere Community Climate Model as its atmospheric component (CESM2-WACCM6, hereafter CESM2-WACCM or CESM2-W; Richter et al., 2021).

Table 1 lists the different S2S forecast systems, and includes relevant information about their hindcast and model configurations. Because the hindcast periods available for different systems vary substantially, unless noted otherwise, we limit our analyses to the 1999-2010 period, which is common to nearly all systems. The one exception is GEFSv12, which has hindcasts that span the period from 2000-2019, for which we simply use the 2000-2010 period. In order to balance including as many prediction systems as possible while also being inclusive of contributions from co-authors with differing capabilities to access and store the large datasets involved, we designated the S2S database systems that provide at least 35-day forecasts as our "core systems", while others were considered optional. Thus, the number of S2S systems shown for a given analysis may vary, but will always include the seven core systems from the S2S database.

There are often multiple model versions of hindcast data available within the S2S database, representing updates to the forecast systems. These model updates sometimes include changes that can reasonably be expected to affect behavior in the stratosphere, such as increases in the number of model levels or lid height, or changes to the atmospheric initialization. Table 1 lists the specific model versions we use; the important ones to note are for the CMA, ECMWF, and ECCC systems. In order to cover the 1999-2010 hindcast period, for ECMWF we consider only CY43R1, CY43R3, and CY45R1 to prevent mixing cycles with large changes to the prediction systems; cycles beyond CY46R1 were excluded because these hindcasts were initialized with ERA5 reanalysis, include updates that explicitly affect mean stratospheric biases (Polichtchouk et al., 2020), and do not fully cover the 1999-2010 period. The CMA and ECCC data both have model versions available in the S2S database that include changes to significantly higher model tops. Since we consider CMA to be a core system (as it performs forecasts beyond 35 days), we use the older low-top version known as the "BCC-CPS-S2Sv1" because the more recent "BCC-CPS-S2Sv2" hindcast period only covers 2004-2018. In contrast, ECCC data is considered to be optional, so some analyses use either the high- or low-top versions; in such cases, we explicitly describe these as "ECCC-hi" or "ECCC-lo", respectively (see Section 2.2 for the specific definitions of high versus low top).

For the analyses herein, we use basic meteorological fields, including zonal and meridional wind components, temperatures, and geopotential heights. These fields from the S2S database are provided once daily at instantaneous 00 UT verification

Table 1. Details of the subseasonal-to-seasonal forecast systems used herein.

Model	S2S Database Version(s)	Hindcast Period	Ensembles	Forecast Span	Model Top
BoM*†	POAMA P24	1981-2013	33	62 days	10 hPa
CESM2-CAM*	–	1999-2019	11	45 days	2 hPa
CESM2-WACCM	–	1999-2019	5	45 days	4.5×10^{-6} hPa
CMA*†	BCC-CPS-S2Sv1	1994-2014	4	60 days	0.5 hPa
CNR-ISAC*	GLOBO	1981-2010	5	32 days	6.8 hPa
CNRM†	CNRM-CM 6.0	1993-2014	15	47 days	0.01 hPa
ECCC-hi	GEPS 6	1998-2017	4	32 days	0.1 hPa
ECCC-lo*	GEPS 4	1995-2014	4	32 days	2 hPa
ECMWF†	CY43R1, CY43R3, and CY45R1	1996-2018	11	46 days	0.01 hPa
GEFSv12*	–	2000-2019	11	35 days	0.2 hPa
GFDL-SPEAR*	–	1991-2020	15	12 months	1 hPa
JMA	GEPS1701	1981-2012	5	34 days	0.01 hPa
KMA†	GloSea5-GC2	1991-2010	3	60 days	85 km
NCEP†	CFSv2	1999-2010	4	44 days	0.02 hPa
UKMO†	GloSea5 and GloSea5-GC2-LI	1993-2016	7	60 days	85 km

CESM2-WACCM hindcast initializations were only performed for September-April

† Core systems

* Systems with low-top models

times on a 1.5° × 1.5 ° latitude/longitude grid, with 10 pressure levels between 1000 and 10 hPa (with about 3 levels in the stratosphere, including 100, 50, and 10 hPa). GEFSv12 fields are provided 6-hourly on a 0.5° × 0.5 ° grid, with 25 pressure levels between 1000 and 1 hPa (6 between 100 and 10 hPa). GFDL-SPEAR fields are also 6-hourly on a 0.5° × 0.5 ° grid, with 33 pressure levels between 1000 and 1 hPa. For the CESM2-CAM and CESM2-WACCM data, we make use of the available "zonal mean" collection of variables that are provided as daily means on pressure levels closest to the model levels; we interpolate this data to a set of 32 standard pressure levels between 1000 and 10 hPa (with 6 between 100 and 10 hPa); these fields have 192 latitudes (∼0.9424° resolution) from the finite volume grid.

The subseasonal hindcast datasets used are all initialized with different atmospheric analyses. Therefore, to ensure that comparisons and biases are all determined with respect to a consistent dataset, we compare hindcast fields to those from the ERA-Interim (ERA-I) reanalysis (Dee et al., 2011). Different modern reanalysis products generally agree very well amongst one another for the time periods (post-1999) and levels (10 hPa and below) we consider (Long et al., 2017; Gerber and Martineau, 2018; Fujiwara, M. et al., 2021), and thus our results are unlikely to be sensitive to the choice of reanalysis.

 ## 2.2 Methods

Throughout the paper we distinguish between forecast systems with low- and high-top atmospheric models. We define the systems having model tops at or above 0.1 hPa with several levels above 1 hPa as high-top; any others that do not meet these criteria are specified as low-top. In total, we consider 8 systems with high-top models and 7 systems with low-top models (see Table 1); however, the precise number of models included in a given analysis varies. In our figures, we highlight low-top models with asterisks and/or dashed lines.

For a given diagnostic, raw biases among the forecast systems are computed by taking the difference between the ensemble-mean hindcasts and ERA-I. We generally composite these biases according to lead time and/or season in order to determine the systematic differences of the hindcast predictions from reanalysis. For some applications we derive lead-time dependent climatologies for the different forecast systems, which we use to determine forecast anomalies and to apply bias-correction. For systems in the S2S database, these climatologies are found by averaging all ensemble-mean hindcasts for a given day of year at a specific lead time. For systems that provide a fixed set of hindcast initializations that do not uniformly cover the same days of year in the hindcasts (such as the GEFSv12 and the CESM2 systems), we create the lead-dependent climatologies according to the method outlined in Pegion et al. (2019); briefly, this method involves averaging all hindcasts for a given day of year and lead time (which is generally less than the total number of years in the hindcast archive), and then applying a rolling 31-day average with centered triangular weights to the "raw" and noisy lead-dependent climatology. Hindcast anomalies are then determined by subtracting these climatologies for a given day of year and lead time from the raw forecast quantities. In some cases we apply a linear bias-correction to raw quantities in order to remove any climatological drift that may exist. This is done by removing the difference between each system's lead time-dependent climatology and the ERA-I climatology from the predicted quantity in question: $Q_{BC}(t,l) = Q_{Raw}(t,l) - [\bar{Q}_{hc}(t_{doy},l) - \bar{Q}_{obs}(t_{doy}+l)]$, where Q_{BC} is the bias corrected quantity, Q_{Raw} is the raw quantity, $\bar{Q}_{hc}$ is the hindcast climatology, $\bar{Q}_{obs}$ is the observed/reanalysis climatology, t is the forecast initialization date, l is the lead time, and t_{doy} is the day of year of the initialization date t.

In sections 3.3 we investigate whether there are biases in relation to extreme stratospheric events, including SSWs and strong vortex events in the NH. SSWs are defined using ERA-I data, based on reversals in the 10 hPa 60°N zonal mean zonal winds, consistent with Charlton and Polvani (2007) and Butler et al. (2017). Similarly to SSWs, strong vortex events are defined using ERA-I data when the daily mean 10 hPa 60°N zonal mean zonal winds exceed 41.2 m/s for at least two consecutive days, corresponding to the 80th percentile of the November-March 1980-2012 zonal winds, consistent with the definition used by Tripathi et al. (2015). These definitions have also been employed in prior investigations of the stratosphere in S2S systems (Domeisen et al., 2020a, b). The central dates of these events are considered to be the first day on which these zonal wind thresholds are met.

3 Results

3.1 Global and Zonal Mean Biases

From the perspective of model evaluation, examining mean state biases in the stratosphere is useful since the processes that govern the distribution of zonal mean temperatures and winds in the stratosphere are quite well understood. Global- and annual-mean temperatures in the stratosphere should be very close to radiative equilibrium (e.g., Olaguer et al., 1992). In contrast, seasonal and meridional variations in zonal mean stratospheric temperatures and winds arise from dynamic influences such as wave-mean flow interaction. Meridional circulations driven by the dissipation of waves drive stratospheric temperatures away from radiative equilibrium, which affects the zonal-mean zonal winds through thermal-wind balance (e.g., Andrews et al., 1987; Shepherd, 2000). Model biases in global annual-mean stratospheric temperatures, and/or seasonally-varying biases in zonal mean temperatures and winds therefore provide information about the likely origin of model errors, either pointing to model components that affect radiative processes, or those that affect middle atmosphere dynamical processes such as parameterized gravity wave drag.

We first consider the annual, global mean temperature biases among the S2S forecast systems. Figure 1 shows these biases for 10, 50, and 100 hPa for each of the models at a lead time of 4 weeks (days 22-28). These biases generally develop and increase in magnitude monotonically with lead time, with many of these biases being present at earlier lead times of 1-2 weeks (not shown). The magnitude of biases tend to be largest in the middle stratosphere at 10 hPa, with 6 of the models shown exceeding absolute biases over 2 K. Positive temperature biases tend to be most common across the models and levels, with the exception of ECMWF and CESM2-WACCM, which primarily have negative temperature biases.

There are some apparent differences between the annual, global mean stratospheric temperature biases between the systems with high- versus low-top models (denoted with asterisks). Warm biases are more common across the pressure levels in the low-top systems, and the highest magnitude biases are generally at 10 hPa, which is likely related to the model tops being relatively close to this level. Some of the low-top systems (CMA-S2Sv1, and CNR-ISAC) have biases that are much less severe in the lower stratosphere and more comparable to the high-top systems. The biases for the high-top systems are generally smaller in magnitude, but there are some exceptions: for instance, the CESM2-WACCM, ECMWF, and NCEP systems have biases at some levels that are as large or larger in magnitude than the low-top systems. Annual-mean global-mean temperature differences between the high- and low-top composites are thus only significant at 50 hPa (numbers in bold), where the low-top systems have only positive biases, and the high-top systems have mostly slight negative biases. However, it is still worth noting that at 10 hPa the magnitude of the low-top biases all exceed 1 K, while all high-top systems except for NCEP are below 1 K.

Figure 1 also highlights some "familial" relationships . In particular, the "NOAA-family" of forecast systems (including GEFSv12, GFDL-SPEAR, and NCEP) all show global mean warm biases throughout the lower and middle stratosphere. The biases in the KMA and UKMO systems are very similar, likely owing to the fact they both use the same GloSea-5 model (see Table 1). There are large differences between the biases apparent in CESM2-CAM and CESM2-WACCM, which may be due to a combination of factors: aside from the differences in model tops, CESM2-WACCM includes fully interactive tropospheric
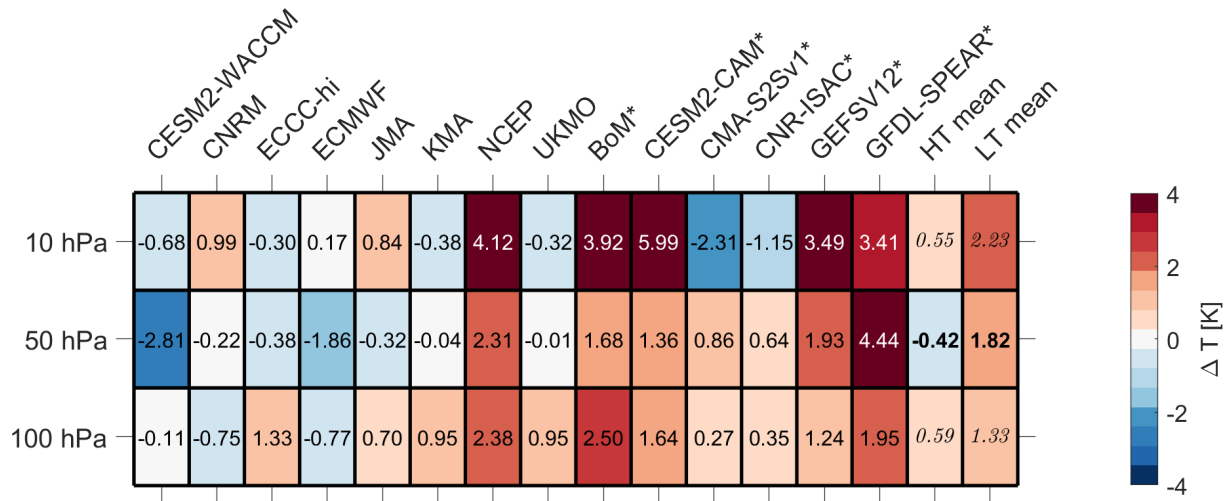
Figure 1. Global- and annual-mean temperature biases at a lead time of 4 weeks. The columns show individual forecast systems, while the rows correspond to different pressure levels. The numerical values displayed are shown in Kelvin. Asterisks at the end of names denote the systems with low-top models. The last two columns show the high top (HT) and low top (LT) composites; the bolded/italicized text in these boxes denote composite differences that are/are not significant at the 95% level from a Student's t-test.

and stratospheric chemistry (Gettelman et al., 2019; Richter et al., 2021), and the atmospheric models are initialized with different reanalyses ("CFSv2" for CESM2-CAM, and "MERRA-2" for CESM2-WACCM).

215   The zonal mean biases across the models are broadly consistent with the annual global mean temperature biases shown above, but reveal further important details about their vertical and meridional structures. Figure 2 shows the zonal mean biases and mean absolute errors (MAE) in temperatures and winds as a function of latitude and height for different seasons, composited for high- and low-top models at a lead time of 4 weeks. These biases and errors are shown for individual systems in the supplemental figures. We note that the interpretation of vertical variations in Figure 2 and the supplemental figures

220   requires some caution since the data from the S2S database are only provided on roughly 3 levels in the stratosphere (100, 50, and 10 hPa). Furthermore, since the UKMO and KMA systems both make use of the GloSea-5 model with the same atmospheric configurations, KMA has been left out of the high top composite in Figure 2 so as not to unfairly weight the high-top composite; as the supplemental Figures S12 and S14 show, these systems have nearly identical biases. The patterns of biases in temperatures and zonal winds are generally consistent across the high- and low-top composites with signatures of (1) global

225   mean warm biases in the stratosphere (consistent with Figure 1), (2) cold extratropical UTLS biases in both hemispheres, (3) easterly wind biases in the tropical stratosphere, and (4) too strong/cold stratospheric polar vortices in the winter hemispheres. It is clear, however, that the biases and MAE in the low-top systems are generally much larger in magnitude despite the similarities in spatial patterns.

       The cold extratropical UTLS and cold winter pole/strong polar vortex biases are recognized as longstanding issues in forecast

230   and climate models. The former issue is generally thought to be related to excessive longwave cooling from a moist bias that
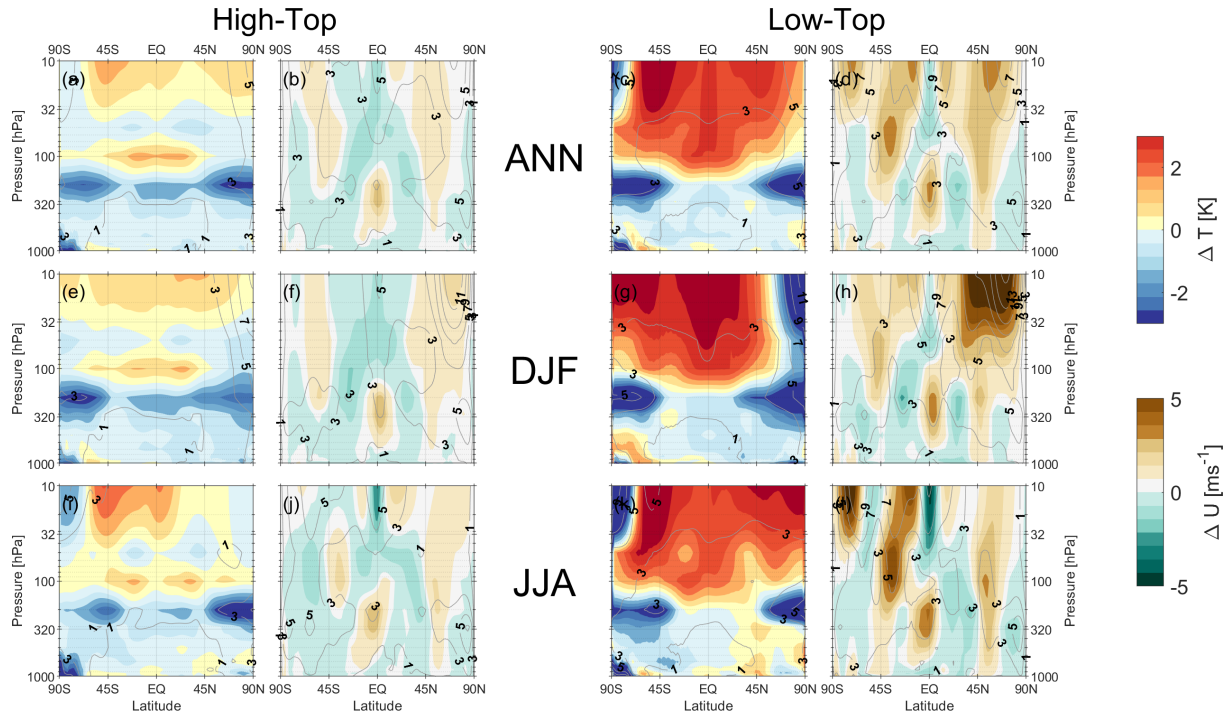
Figure 2. Zonal mean (a,c,e,g,i,k) temperature and (b,d,f,h,j,l) zonal wind biases as a function of pressure and latitude at a lead time of 4 weeks. Data are averaged over (a-d) the full annual cycle, (e-h) boreal winter (DJF) and (f-l) austral winter (JJA), and are shown separately for (a,b,e,f,i,j) high-top and (c,d,g,h,k,l) low-top models. The grey line contours show the composites of mean absolute errors from each system.

is present in initial conditions, and/or develops over time due to an inability to properly maintain the distribution of water vapor in the region of the tropopause (see, e.g., Bland et al., 2021, and references therein). Figure 2 and the supplemental figures show that these cold biases are still apparent in high-top models, but are generally smaller in magnitude. For the SH summer, the difference in the cold UTLS temperature bias between the high- and low-top systems is significant at 200 hPa (Figure S15c). These differences may be related to the fact that the systems with high-top models generally have more model levels/higher vertical resolution, making them better able to represent processes in the tropopause region. The too strong/cold winter stratospheric polar vortex biases, particularly in the NH, reflect a lack of dynamical variability. Figure 2 shows that the low-top systems have higher MAE, more pronounced cold winter poles, and stronger polar vortex winds in both hemispheres; for the NH winter, the differences between the high- and low-top systems are significant at 10 hPa (Figure S15c and d). These results are consistent with prior studies that find models with tops below the stratopause generally fail to realistically simulate stratospheric variability (e.g., Charlton-Perez et al., 2013; Shaw et al., 2014; Rao and Garfinkel, 2021a). There are multiple reasons why these biases are generally worse in low-top models, usually related to an underestimation of

resolved/parameterized gravity wave drag (see e.g., Section 3.3), as well as possible unphysical effects of the model lid (e.g., Shaw and Perlwitz, 2010; Richter et al., 2014).

245   The apparent easterly bias in the tropical stratospheric zonal winds across the high- and low-top systems represents a general bias among the models in representing the QBO. We examine these biases in the next section in more detail.

3.2   Biases Related to the Tropical Stratosphere and QBO

The circulation of the tropical stratosphere is dominated by the QBO, and hence the zonal mean biases in the tropics shown above must be partially related to each forecast system's ability to maintain the QBO from initial conditions. To the extent
250   that S2S models can represent the QBO and its teleconnections accurately, the QBO could lead to more reliable predictions on subseasonal and seasonal timescales in the troposphere (Garfinkel et al., 2018; Merryfield et al., 2020). Therefore, we now consider whether this potential is realized by the S2S forecast systems. Unlike elsewhere in the paper, here our QBO analyses make use of the full hindcast periods available to each model to maximize the number of QBO cycles available when determining biases (see Table 1).

255   Figure 3 considers whether these models are capable of maintaining the anomalous QBO winds present in initializations from November through February (NDJF). In the top row, we composite hindcasts in which the zonal wind anomalies (computed with respect to each model's climatology) at 10 hPa, 5°S-5°N averaged over the first three days of the hindcast exceed 3 m/s as being westerly QBO (WQBO), and less than -3 m/s as easterly QBO (EQBO); we subsample the ERA-I reanalysis anomalies to match the dates included in each model's composite before calculating differences. The middle row shows similar
260   quantities, but with the QBO phase composites determined using the zonal wind anomalies at 50 hPa instead. On subseasonal timescales out to 8 weeks, the reanalysis quantities for the QBO remain virtually constant (see supplement Figure S16). At 10 hPa (Fig 3, top row), systems with low-top models (dashed lines) clearly struggle to maintain the amplitude of both EQBO and WQBO initialized anomalies, which start out biased in week 1 by roughly 5 m/s, and decay with lead time compared to reanalysis by an additional 5 m/s for EQBO and more for WQBO. The net effect is that the QBO signal ($\pm$10-20 m/s;
265   Figure S16a and b) is almost entirely lost for some of the low-top systems. The systems with high-top models generally show much slower and slighter decay, with relatively small anomaly differences from reanalysis close to 0 m/s across lead times. At 50 hPa the reanalysis winds generally range from $\pm$10-15 m/s (Figure S16c and d); in the forecasts (Figure 3, middle row), even the high-top systems show an apparent decay of anomalies with lead time. The differences from reanalysis are slightly more apparent in the 50 hPa EQBO composite, for which a larger fraction of the model wind anomalies differ from reanalysis
270   by 5 m/s or more at the end of the forecasts compared to WQBO. The main exceptions are the UKMO and KMA systems, which share the same atmospheric models, and perform well relative to the other systems as demonstrated by their reanalysis differences staying close to 0 m/s. Overall, the fact that the subseasonal forecast systems are generally better able to simulate and maintain the QBO in the middle stratosphere versus the lower stratosphere is similar to what has been found in climate models, seasonal forecast systems, and models participating in the "QBO-initiative" (QBOi; Richter et al., 2020; Rao et al.,
275   2020a; Bushell et al., 2020; Stockdale et al., 2020). This difference in success between the mid- and lower-stratospheric QBO could be related to several factors, including gravity wave parameterizations, the need for high vertical resolution in the lower
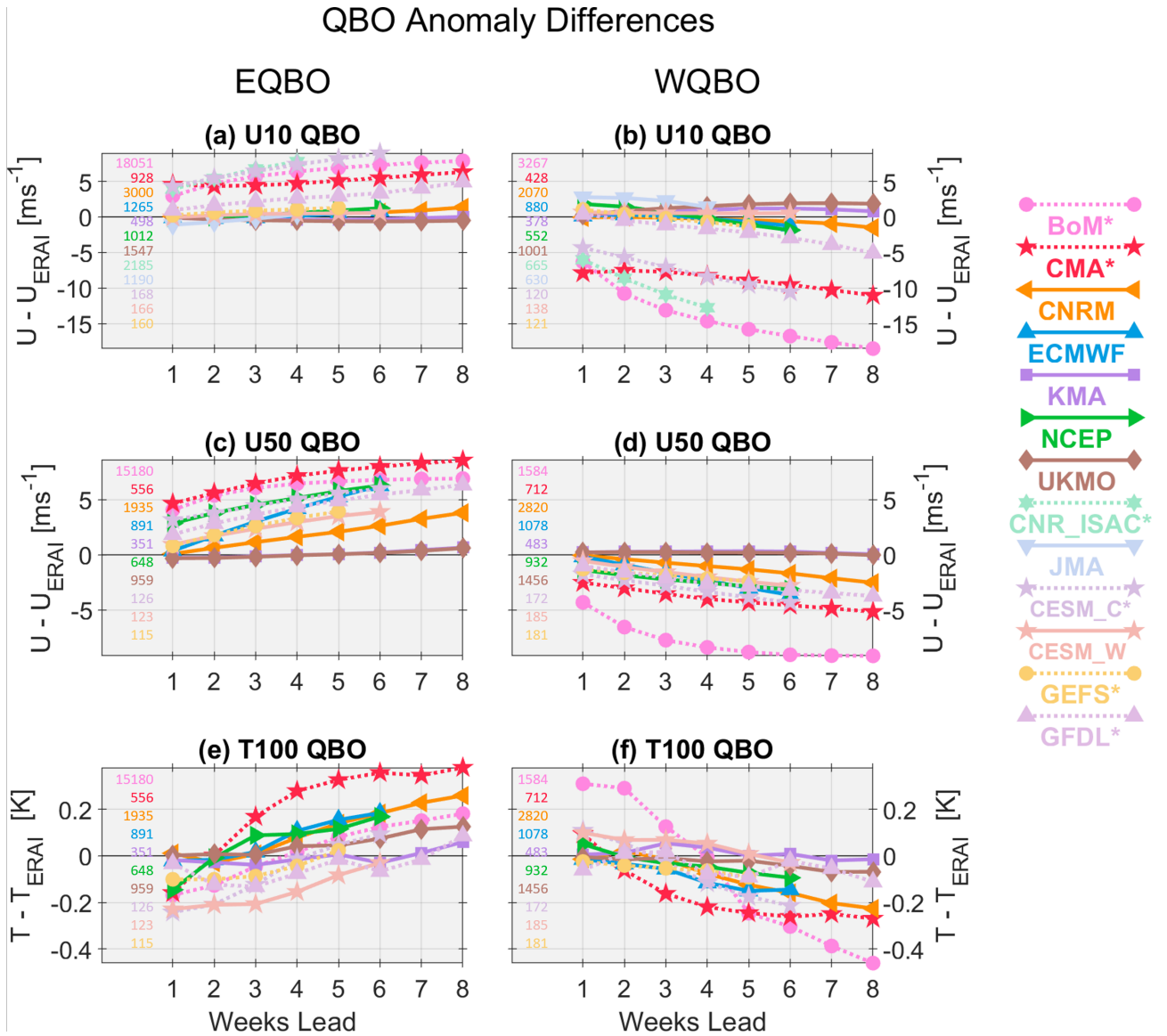
Figure 3. Time series of tropical stratospheric anomaly differences (from reanalysis) composited for easterly and westerly phases of the QBO in the subseasonal hindcasts. (top row) QBO defined using 10 hPa, 5°S-5°N zonal wind; (middle row) QBO defined using 50 hPa, 5°S-5°N zonal wind; and (bottom row) 100 hPa, 5°S-5°N temperature in the QBO phases defined using 50 hPa winds. All composites are based on November-February initializations only, with the number of initializations in each composite shown in each panel.

stratosphere to realistically capture upward wave flux and subsequent downward QBO propagation, and possible influences from model vertical diffusion in the lower stratosphere (Geller et al., 2016; Garfinkel et al., 2021; Polichtchouk et al., 2021).

The QBO has been shown to influence tropical convection on subseasonal timescales, and one of the leading mechanisms for this effect is related to the QBO's mean meridional circulation, which leads to temperature (and buoyancy frequency) anomalies in the tropical tropopause layer that subsequently affect high clouds and convection (Gray et al., 2018). These relationships also underpin the QBO's observed relationship with the Madden-Julian Oscillation (MJO; Yoo and Son, 2016; Son et al., 2017; Lee and Klingaman, 2018; Martin et al., 2021a), which can modulate MJO forecast skill (Kim et al., 2019; Lim et al., 2019). To this end, we examine the associated temperature anomalies in the lowermost tropical stratosphere at 100 hPa (T100; bottom row of Figure 3). These biases are composited based on the initial QBO wind anomalies at 50 hPa (consistent with the middle row). The temperature anomalies are proportional to the QBO-related shear in the lower stratosphere with typical EQBO/WQBO wind anomalies corresponding to cold/warm anomalies of ±0.5 K, as determined from reanalysis (Figure S16e and f). These temperature anomalies are present in the initialized states of the models, with some systems overestimating the magnitude of the anomalies at week 1. However, in most of the systems, these T100 anomalies decay with leadtime and eventually switch sign such that the models become warmer relative to reanalysis in EQBO, and colder in WQBO. The rate of weakening differs among the models, with the low-top systems generally showing much more rapid decay of anomalies. There appears to be little difference between the EQBO and WQBO composites in the ability of models to maintain the tropical lower stratosphere temperature anomalies. Underestimating the amplitude of the QBO in the lower stratosphere (both in winds and temperatures) down to the tropopause is an issue similar to that present in CMIP6 and QBOi models (Bushell et al., 2020; Richter et al., 2020), and may be a contributing factor to why many subseasonal forecast systems show insignificant relationships between the QBO and MJO (e.g., Kim et al., 2019).

The QBO is known to have an important teleconnection to the boreal winter polar vortex known as the "Holton-Tan effect" (e.g., Holton and Tan, 1980; Baldwin et al., 2001; Calvo et al., 2007; Garfinkel et al., 2012; Anstey and Shepherd, 2014), in which weaker or stronger polar vortex winds preferentially occur with EQBO or WQBO winds in the lower stratosphere, respectively. Figure 4 shows the composite anomalies in 10 hPa polar vortex winds, and 100 hPa polar cap geopotential heights (Z100) composited based on the 50 hPa QBO phase at initialization. Since the Holton-Tan effect develops in early winter and is most pronounced in mid-winter, Figure 4 is limited to forecast initializations within November and December, before the effect is strongly embedded in initial conditions. In the reanalysis EQBO composites, the weak vortex signal becomes apparent beyond roughly 3 weeks, with easterly wind anomalies on the order of 5-10 m/s that are maintained at longer lead times (Figure S17a). While most of the subseasonal forecast systems track reanalysis closely out to 3-4 weeks, it is clear that they underestimate the magnitude of polar vortex weakening (Fig 4a), with winds that are too westerly at longer leads. The results for WQBO (Fig 4b and Fig S17b) are largely similar, with the subseasonal forecasts having an easterly wind bias at longer leads. The bottom row of Figure 4 shows the anomalies in polar cap geopotential heights at 100 hPa, where persistent anomalies in the strength of the lower stratospheric vortex are more closely tied to surface-related impacts. Here the results are similar to the 10 hPa polar vortex winds, with most systems failing to match the change in amplitude of the anomalies with lead time, which can exceed ±40-60m in reanalysis (Figure S17c and d).

Figure 4. As in Figure 3, but for polar stratospheric quantities in QBO composites based on the initial 50 hPa winds from November and December initializations, including (top row) zonal mean zonal wind at 10h Pa, 60°N, and (bottom row) 100hPa, 70°-90°N polar cap geopotential heights (Z100). The number of initializations in the composites are shown on each panel.

3.3 Northern Hemisphere Polar Vortex Variability

Variability in the NH polar winter stratosphere primarily arises from extreme dynamical polar vortex events, including midwinter SSWs and strong vortex events. The occurrence of these events are generally associated with extremes in the upward wave fluxes that disturb the polar vortex, with SSWs/strong vortex events being preceded by extended periods of above/below-

315

normal wave driving, respectively (e.g., Polvani and Waugh, 2004). A typical metric for upward planetary wave flux is the meridional eddy heat flux, $v'T'$ (with v being the meridional wind, T the temperature, and the primes denoting deviations from the zonal mean). In models, such wave driving should be well represented, dependent upon tropospheric variability and proper simulation of vertical wave propagation. There can also be significant variability among polar vortex events in terms of

320 different characteristics, such as their timing, magnitude, persistence, and even polar vortex geometry (e.g., Karpechko et al., 2017). Importantly, the occurrence of such polar vortex events can lead to coupling with the troposphere that lasts for weeks to months; in forecast models, the occurrence of these events can improve tropospheric predictability by providing "forecast windows of opportunity" (e.g., Butler et al., 2019a, and references therein).

We first examine and compare biases in the NH eddy heat fluxes among the subseasonal forecast systems. Figure 5 shows

325 the December-February eddy heat flux biases with respect to ERA-I at a lead time of 4 weeks. The climatological DJF 100 hPa heat flux from ERA-I is shown in Figure 5a, while the high- and low-top composite biases are shown in b and c. The observed climatological heat fluxes show two centers of action, with one over the North Pacific, and another over Scandinavia/Siberia. This pattern largely represents the influence of planetary-scale zonal waves 1 and 2, which generally have the most impact on the stratosphere. The mean biases in the subseasonal forecast systems strongly differ between the high- versus low-top

330 composites. The high-top systems underestimate heat fluxes in the Pacific region more than over Scandinavia/Siberia, which manifests as a slight negative bias in wave-1 heat fluxes (Fig 5d), but none of the biases are statistically significant. In contrast, the low-top systems significantly underestimate heat fluxes in both regions, while also overestimating heat fluxes over Canada and Greenland; the latter likely indicates the low-top systems do not capture the region of negative heat fluxes seen in the ERA-I climatology, a region commonly influenced by downward wave reflection (Matthias and Kretschmer, 2020; Cohen et al., 2021;

335 Messori et al., 2022).

These biases in regional heat fluxes are indicative of biases in the heat flux contributions from zonal waves 1 and 2, and thus we show the time evolution of these in Figure 5d and e. In most of the models, the week-1 wave-1 heat flux biases are small in magnitude, except for low-top systems such as CMA and BoM. As mentioned above, the greater underestimation of heat fluxes over the Pacific in the high-top models is indicative of a slight negative wave-1 bias, which is most apparent in the

340 CNRM and ECMWF systems. However, the low-top systems show negative biases that are much larger in magnitude from week 3 and beyond, especially in the BoM and CMA systems. Here the NCEP system stands out since it actually overestimates the wave-1 heat flux for weeks 3-5. The results are similar for the wave-2 heat flux biases in Figure 5e, which shows the low-top systems more strongly underestimate the heat fluxes across lead times compared to the high-top systems. The high-top systems all show very small wave-2 heat flux biases out to about week 6, after which the CNRM, KMA, and UKMO systems

345 have a slight positive bias. Overall these results reveal that the low-top systems consistently underestimate the contributions of planetary-scale waves to eddy heat fluxes in the lower stratosphere, which is consistent with the strong vortex/cold pole bias shown in the low-top zonal mean composite from Figure 2.

The biases in the stratospheric background circulation (section 3.1) and heat fluxes described above can affect the occurrence and timing of threshold events such as SSWs and strong vortex events. Figure 6 shows the probability of occurrence of 10 hPa

350 60°N winds less than 0 m/s (corresponding to SSWs) or greater than 41.2 m/s (corresponding to strong vortex events) for
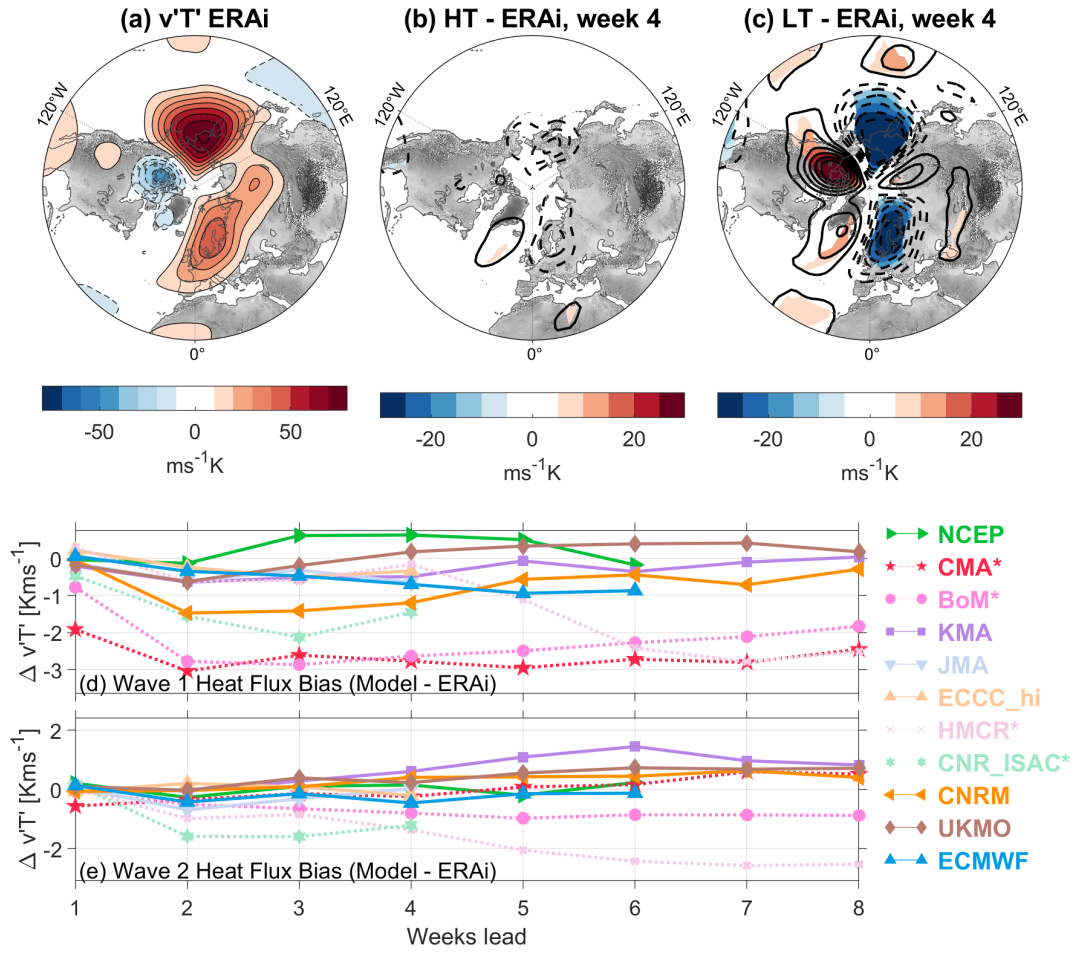
Figure 5. (a) The December-February climatology of ERA-I eddy heat fluxes, $v'T'$, at 100 hPa. (b) The high-top composite of week-4 eddy heat flux biases with respect to ERA-I from November-January initializations. (c) is as in (b), but composited for the low-top models. In panel a, the line and colour-filled contours match the colorbar spacing of 10 ms^{-1}K; in panels b and c, the line contours match the colorbar contour intervals of 5 ms^{-1}K, but colours are only shown where the biases are statistically significant at the 95% level from a two-tailed Student's t-test. (d) Time series of the difference between the S2S hindcasts and ERA-I for the wave-1 eddy heat flux averaged over 45°-75°N for lead times from 1 to 8 weeks. (e) is as in (d), but for wave-2.

different weekly lead times. These probabilities are composited based on initialization, so for instance, the week 4 values of the January initializations (Figure 6e,f) include verification times from the month of February. The seasonal distributions from reanalysis (colored circles) indicate low probabilities of SSW events in early winter (November and December), with the highest occurrence of events being in late winter (January-March). Note that since the figure is composited based on initializations, the probabilities for March include verification times in April, and so the corresponding probabilities for easterly winds are likely influenced by final warmings. Regardless, this seasonal cycle is only partially reproduced in the subseasonal

models, which particularly underestimate the probability of events for December and January initializations at weeks 3 and 4, and overestimate the probability for March initializations at weeks 3 and 4. This bias is in agreement with results from climate models (Ayarzagüena et al., 2020; Tyrrell et al., 2021), which tend to exhibit a peak in SSW occurrence in late winter instead of in January, and seasonal prediction models, which also fail to reproduce the SSW peak in January (Portal et al., 2021) despite the seasonal average of SSWs often being well reproduced (Domeisen et al., 2015). Interestingly, the NCEP system consistently predicts a higher occurrence of easterly winds than other systems. This means the NCEP system is more accurate for week 4 SSW risk forecasts initialized in December and January, but then overestimates SSW probabilities in February and March. The NCEP system's higher prediction of easterly winds is possibly related to its significant weak vortex bias (see Supplement Figure S13), which may be linked to its overestimation of heat fluxes (Figure 5d).

Strong vortex events (right column of Figure 6) exhibit an entirely different seasonal cycle, with most events occurring between December and January (primarily due to the threshold-based definition and the climatological maximum strength of the vortex occurring in these months). Most models tend to underestimate the frequency of strong polar vortex winds for November and December initializations, particularly at weeks 3 and 4. There are some notable exceptions, including the CNR-ISAC, CNRM, GEFSv12, and CESM2-CAM systems, which all have substantial strong polar vortex biases in their wintertime zonal winds (Supplemental Figures 2, 5, 6, and 9). For January initializations, most models instead overestimate the frequency of strong vortex winds beyond week 1 (Fig. 6f), which is also consistent with the general strong vortex biases evident in the model composites of Figure 2.

In addition to the probability of the events from raw forecasts, the black horizontal lines in Figure 6 indicate the probabilities estimated from bias-corrected forecasts (see Section 2.2 for details of the mean bias-correction process). The probability of both SSWs and strong vortex events in the bias-corrected hindcasts initialised in November, December and January is generally either close to, or smaller than, the observed probabilities from ERA-I across all forecast systems. In most cases, this corresponds to an improvement over the raw forecasts. Especially for the prediction of SSWs for forecasts initialized in January (Fig. 6e), the mean bias correction clearly improves the estimates over those from the raw data, particularly for lead times of 3-4 weeks. However, the bias correction for late winter/early spring predictions (initializations in February and March, weeks 3 and 4) does not necessarily bring the easterly wind probabilities closer to observations. In some cases the bias correction increases the probability of events, even for systems whose un-corrected probabilities already closely match reanalysis. This may mean that model zonal-mean zonal wind biases in late winter and early spring tend to not dynamically alter the probability of zonal wind reversals at times when final warmings may be expected to occur. The exceptions here are the systems with the most severe biases, such as BoM and CNR-ISAC. Nevertheless, and especially for early winter, the magnitude of zonal wind biases clearly changes the probability of fixed threshold events in most of the S2S systems. This supports the utility of bias correction for stratospheric S2S forecasts, albeit with some limitations. Furthermore, such bias correction has to be applied and interpreted with care, since the non-linear dynamics in the models evolve according to their own potentially biased mean states, and therefore, little can be said about potential tropospheric responses to such bias-corrected stratospheric forecasts.

There also exist biases in the magnitude of predicted events, even at relatively short lead times. In Figure 6 we identified the probability of polar vortex events occurring in the subseasonal forecasts at different lead times; in Figure 7 we instead focus
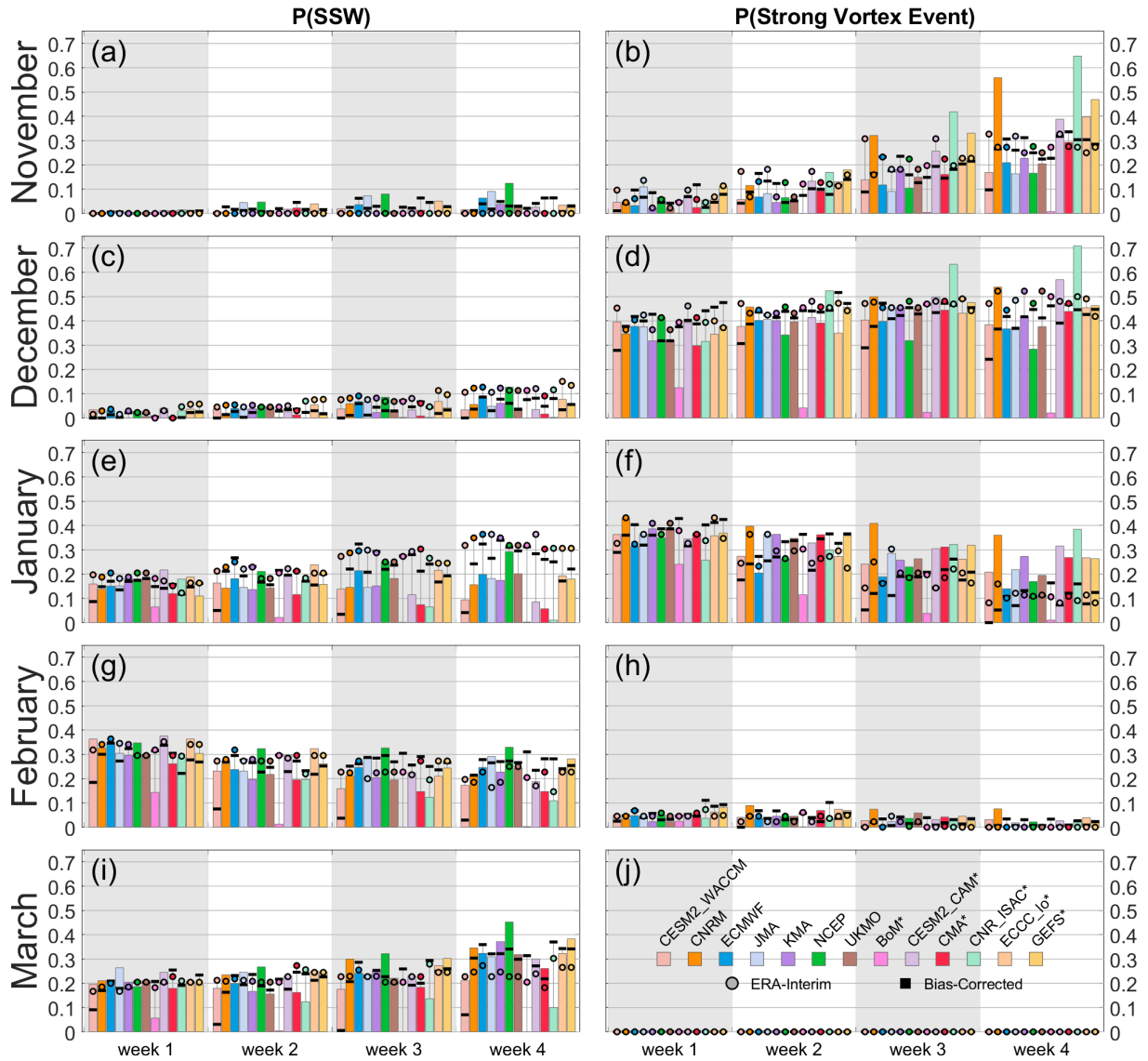
Figure 6. Probability of boreal (left column) sudden stratospheric warmings and (right column) strong vortex events, shown individually for composites of initializations within each month (rows) and weekly lead time (horizontal axis, alternating grey/white background shading within panels). For each model, the solid bars show the raw estimates, while the bold black horizontal lines indicate the probability determined after bias-correction. The coloured circles indicate the probabilities computed using ERA-I subsampled to match the same dates from each individual set of model hindcasts.

on observed polar vortex events in the ERA-I record, and assess the forecasted wind changes at verification times surrounding the observed events. Figure 7 shows the distributions of simulated wind changes associated with SSWs and strong vortex events (defined as in Section 2.2) in the 1999-2010 reanalysis record. The deceleration or acceleration associated with these

395 events is measured by computing the change in the hindcast/reanalysis zonal-mean zonal wind at 10 hPa and 60°N, at ± 5 days around the ERA-I event onset dates; for the hindcasts, these are first computed individually for each ensemble member before being composited. Almost all systems underestimate the wind changes for both SSW and strong vortex events at all lead times, yielding dominantly positive/negative mean errors for SSWs/strong vortex events, respectively. However, the prediction of the observed magnitude of events clearly improves with decreasing lead time, as expected. At weeks 3 and 4 lead times,

400 the predicted wind change distributions among the systems are generally close to zero and exhibit small spread. This indicates that these models predict climatological zonal-mean zonal wind values or only weak wind tendencies of the same sign as the events. This is not a shortcoming of the prediction systems, but rather to be expected given that the typical predictability limit for these vortex events is about two weeks. Even within a lead time of two weeks, some systems still underestimate the magnitude and spread of the observed wind changes; many of these are the low-top models such as BoM, CESM2-CAM, and

405 CMA. For SSWs, the ECMWF and NCEP systems have the smallest errors of around 5 m/s; for strong vortex events, the CNRM, CESM2-CAM, and GEFSv12 systems consistently have the lowest mean errors within 10 m/s from 1-3 weeks, but these systems also have substantial positive zonal wind biases. Figure 7c,d also highlights the disparity in the magnitude of wind changes between SSWs and strong vortex events; median errors for SSWs are on the order of 15 - 20 m/s with outliers up to 60 m/s, whereas the range is much smaller for strong vortex events. This reflects the large and sudden deceleration of winds that occur during SSWs that (in absolute terms) is much larger than the acceleration of winds for strong events.

410 that occur during SSWs that (in absolute terms) is much larger than the acceleration of winds for strong events.

Finally, we examine whether there are biases among the subseasonal models in forecasting the geometry of the polar vortex at times surrounding observed SSW events. The shape and location of the polar vortex are ultimately affected by the vertical wave activity that influences the occurrence and/or magnitude of SSWs. We examine vortex geometry using elliptical diagnostics, which provide quantities such as the vortex centroid latitude and aspect ratio (e.g., Waugh, 1997; Seviour et al., 2013) that can

415 be used to quantify the displacement or stretch of the vortex during SSWs. We perform these calculations using the hindcast 10 hPa geopotential heights, assuming that the 30 km contour is representative of the vortex edge. Figure 8 shows the ensemble-mean predictions of the centroid latitude and aspect ratio diagnostics as a function of lead time and initialization with respect to composites of the central dates of displacement and split SSWs within 1999-2010. Note that the colorbars of Figure 8 transition to pink colors at 66°N for centroid latitude, and 2.4 for aspect ratio, corresponding to the thresholds used in Seviour et al. (2013)

420 to define displacement and split SSWs. For displacement events (Fig 8a-i), most models capture a latitudinal deviation from the pole at long lead times of around a month, though with an underestimation of the magnitude of the displacement at lead times beyond 3 weeks. However, the BoM, CMA, and NCEP systems show signs of systematic biases in their predicted centroid latitudes. BoM and CMA (both low-top systems) show virtually no centroid latitude variability at longer lead times, with only the forecasts falling within about two weeks of the SSWs showing significant latitudinal displacements. On the other hand,

425 NCEP appears to have a systematic bias toward a vortex that is too frequently displaced at longer lead times. While Figure 8 is only focused on SSW events, the results for NCEP, BoM, and CMA are consistent with their climatological heat flux biases shown in Figure 5; at longer leads, BoM and CMA consistently underestimate wave-1 heat fluxes, while NCEP consistently overestimates them.
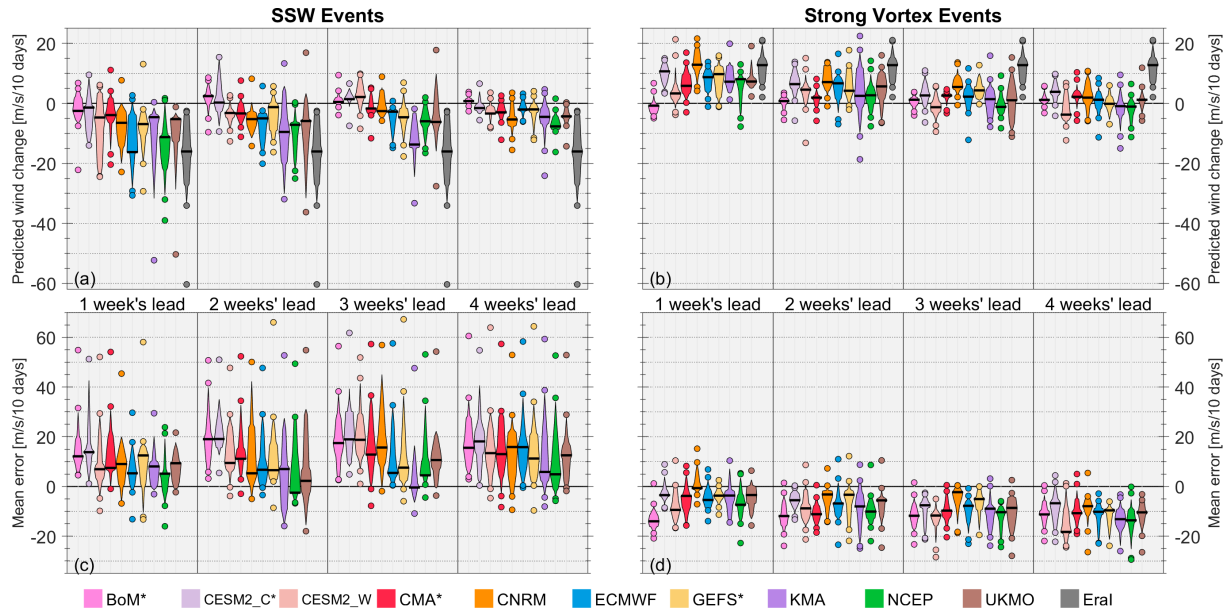
Figure 7. Distributions of the composited wind changes forecasted by each ensemble for forecast verification times surrounding (left column) boreal sudden stratospheric warmings and (right column) strong vortex events, composited over all such events in the 1999-2010 hindcast records. This corresponds to 11 SSW events and 13 strong vortex events across the boreal winters from 1999/2000 to 2009/2010. Units are m/s over the 10-day period centered on the observed event central dates within the forecasts or reanalysis. Within each section of each panel, data are shown as violin plots covering the 15-85% range of wind changes or mean errors, with outliers outside this range indicated with individual coloured circles, and black horizontal lines indicating the median value. In panels (a) and (b) the quantities shown are raw values, while those in panels (c) and (d) are shown as deviations from ERA-I. Note that individual models may contain a slightly different number of samples, especially for longer lead times, due to the different hindcast lengths available for each system (see Table 1); in other words, some hindcasts do not fully cover the 10-day periods surrounding the observed SSWs and strong vortex events, and are excluded.

In the case of vortex split SSWs (Figure 8 j-r), the high-top models perform much better relative to the low-top models, though still worse relative to displacement events. Vortex split events are known to be inherently less predictable than displacement events (e.g., Taguchi, 2018; Domeisen et al., 2020a), but the low-top models only show enhanced aspect ratios within a lead time of roughly 10 days. Of these, the BoM system shows large aspect ratios only in the initializations that are close to the onset dates. While the sample sizes here are quite small with only 6 displacement and 3 split events in the common time period under consideration, these results do show the signatures of the systematic biases among the modeling systems shown previously, particularly those for the wave-1 and wave-2 heat fluxes (Figure 5).

3.4 Southern Hemisphere Polar Vortex Variability

In the Southern Hemisphere (SH), stratospheric polar vortex variability is mainly associated with interannual variability in the timing of the springtime polar vortex break-down (i.e., the final warming) sometime between November and January. Prior to
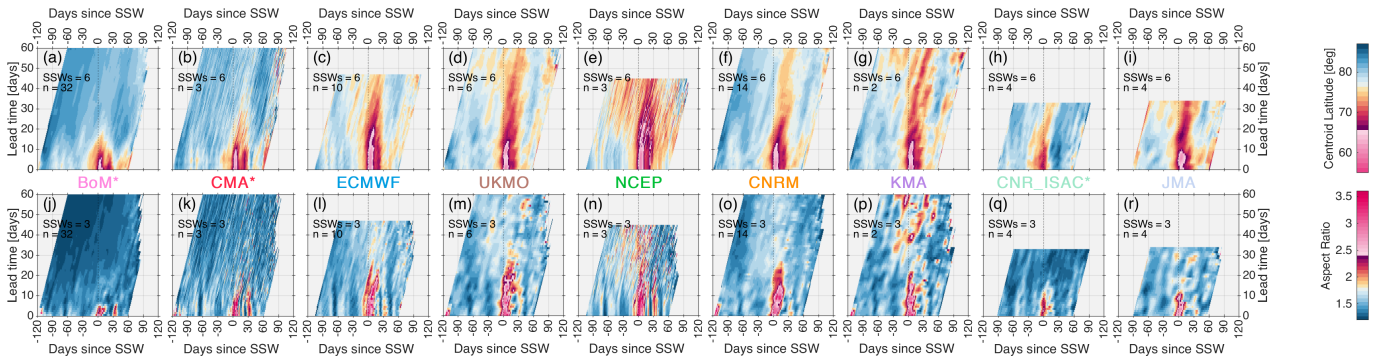
Figure 8. Predictions of boreal stratospheric vortex (a-i) centroid latitude (j-r) aspect ratio, computed at 10 hPa from hindcast geopotential height. Values shown are an average of all ensemble members over all (a-e, f-i) displacement and (j-n, o-r) split SSWs in the 1999-2020 record. Data are plotted against (horizontal axis) SSW-relative day number and (vertical axis) model lead time. For each panel, the number of SSWs ('SSWs') and number of ensemble members ('n') is indicated at top left. The colorbar transitions to pink occur at thresholds of 66°N for centroid latitude, and 2.4 for aspect ratio.

the final warming, the SH polar vortex undergoes a downward shift in its location relative to its midwinter position (Mechoso
et al., 1985). The downward shift of the polar vortex has been linked to a poleward shift of the SH eddy-driven jet in austral
spring, while the timing of the polar vortex breakdown has been linked to the equatorward shift of the eddy-driven jet between
November and January (Hio and Yoden, 2005; Byrne et al., 2017).

Given the above, the SH spring season can be regarded as a 'window of opportunity' for more skillful tropospheric forecasts
on S2S timescales provided that stratospheric variability is accurately represented. Indeed, previous studies have shown that
the SH tropospheric variability during spring, prior to the polar vortex breakdown, can be predicted from stratospheric initial
conditions in winter (Seviour et al., 2014; Lim et al., 2018; Byrne et al., 2019; Rao et al., 2020b; Oh et al., 2022). However,
evaluations of individual seasonal prediction systems such as the ECMWF reveal unrealistic SH stratospheric variability and an
inability to correctly represent stratosphere-troposphere coupling during austral spring, with likely impacts on the tropospheric
mean state during that season (Polichtchouk et al., 2021).

As with the NH, we first explore biases in the wave driving of the SH stratosphere, represented by the 100 hPa eddy heat flux,
$v'T'$ (Figure 9). In the SH, negative eddy heat fluxes are poleward and represent upward propagation of wave activity into the
stratosphere. In ERA-I (Fig. 9a), the poleward eddy heat fluxes are largest over the Southern Ocean with local maxima south
of Australia (150E) and downstream of the Antarctic Peninsula (30W). Eddy heat fluxes climatologically peak in amplitude in
austral spring and are associated primarily with stationary waves of wavenumber-1 with a secondary role from transient waves
(Randel, 1988). Similar to the NH (Figure 5), the spatial patterns of eddy heat flux biases strongly differ between the high-
and low-top composites, particularly at longer lead times. At week 4 (Figure 9c), the low-top composite shows positive biases
over the two local maxima in the ERA-I climatology; these biases are primarily from the BoM and CNR-ISAC systems, which
both have large positive biases over the region between Antarctica and Australia, projecting onto the wave-1 heat fluxes (see
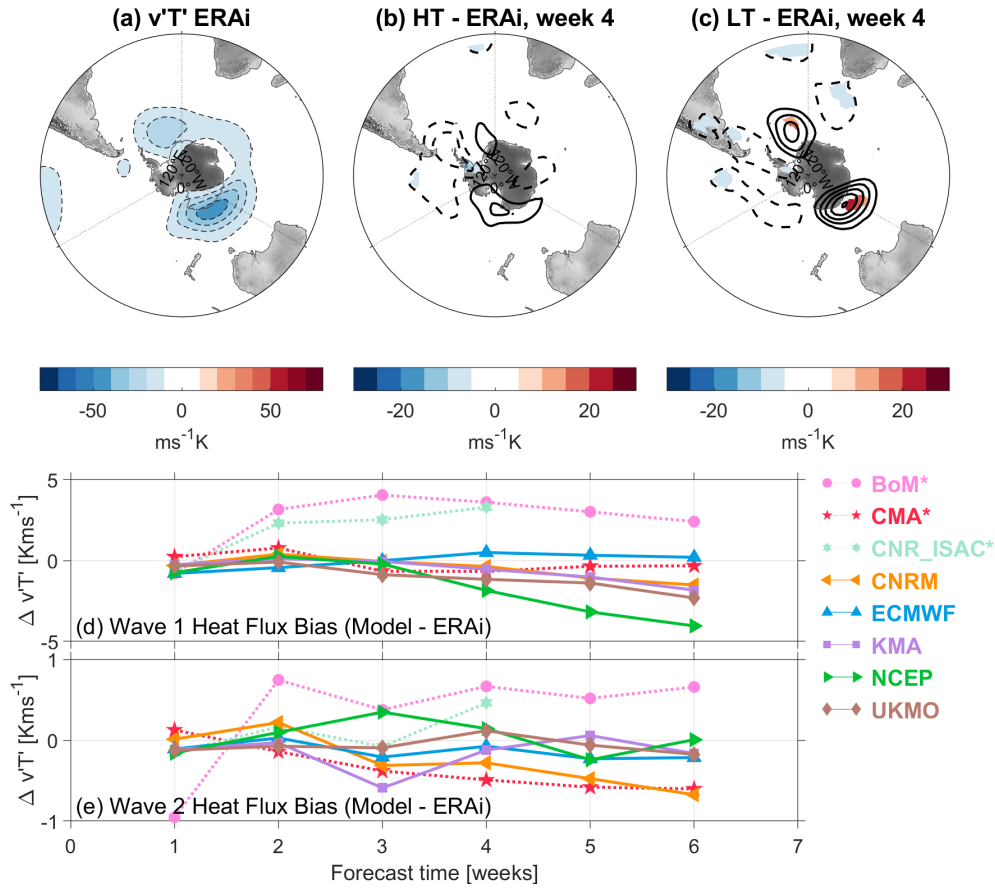
Figure 9. (a) The September-November climatology of ERA-I eddy heat fluxes, $v'T'$, at 100 hPa in the Southern Hemisphere. (b) The high-top composite of eddy heat flux model biases with respect to ERA-I from August-October initializations at a lead time of 4 weeks. (c) is as in (b), but composited for the low-top models. In panel a, the line and colour-filled contours match the colorbar spacing of 10 ms^{-1}K; in panels b and c, the line contours match the colorbar contour intervals of 5 ms^{-1}K, but colours are only shown where the biases are statistically significant at the 95% level from a two-tailed Student's t-test. (d) Time series of the difference between the S2S hindcasts and ERA-I for the combined wave-1 and 2 planetary wave heat flux over 45°-75°S, based on August-October initializations.

also Figure 9d). In contrast, the pattern of biases in the high-top composite is much less coherent except for a relatively large
460 region of negative biases between the southern tip of South America and the Antarctic peninsula. The spatial patterns of SH heat flux biases shown for week 4 are largely consistent with those shown as a function of lead time (Figure 9d and e). Among the low-top systems, BoM and CNR-ISAC show positive biases for both wave-1 and 2 heat fluxes beyond week 1 (indicating decreased poleward eddy heat flux); in contrast, the CMA system has relatively small wave-1 biases, but too negative wave-2 heat fluxes beyond week 1 (indicating enhanced poleward eddy heat flux). Most of the high-top systems also have negative
465 heat flux biases for both wavenumbers at longer leads, especially beyond week 4 for wave-1. Among these, NCEP shows the most negative biases for wave-1 heat fluxes, which nearly reach -5 K m/s at week 6. The fact that the high-top systems seem to
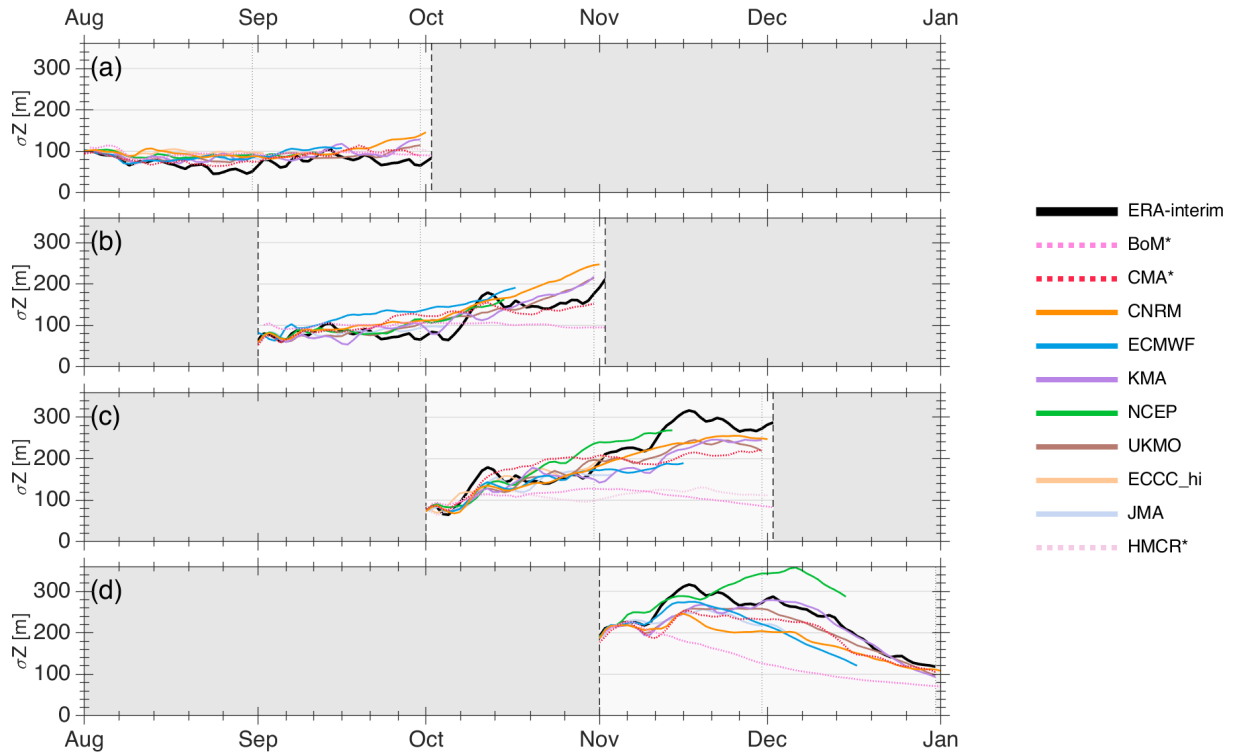
Figure 10. Interannual standard deviation of the southern hemisphere hindcast 50 hPa polar cap (60°-90°S) geopotential height for each model (coloured lines) in comparison to ERA-I (black lines). The hindcasts closest to the first of each month are chosen for each model. The verification time range in each row spans a maximum of 8 weeks.

slightly over-estimate SH upward wave fluxes could imply that these systems also have too high springtime variability in the SH stratosphere.

Following on from the SH eddy heat flux biases, Figure 10 shows the interannual standard deviation of SH polar cap (60°S-

470 90°S) geopotential height at 50 hPa for each of the S2S systems based on initialization dates that are closest to the first of August, September, October, and November. Here we consider the common hindcast period of 1999-2010, but exclude 2002, the year of the only major sudden stratospheric warming in the SH. In ERA-I, the observed interannual variability increases considerably from the beginning of October, peaking in mid-November and decaying thereafter. The hindcasts initialized in August and September present similar levels of variability as reanalysis up to early October; for initializations in early Septem-

475 ber, several models (CNRM, ECMWF, KMA, UKMO) even show an increase in variability at longer lead times. However, the difference between the models and observations increases for initializations in early October and November, with most models underestimating the observed variability beyond 4 weeks. For early November initializations, most of the high top models (with the exception of NCEP) underestimate the variability at shorter lead times from roughly week 2 onwards. Clearly though, the low-top models show the most consistent underestimation across the different months and lead times; BoM and
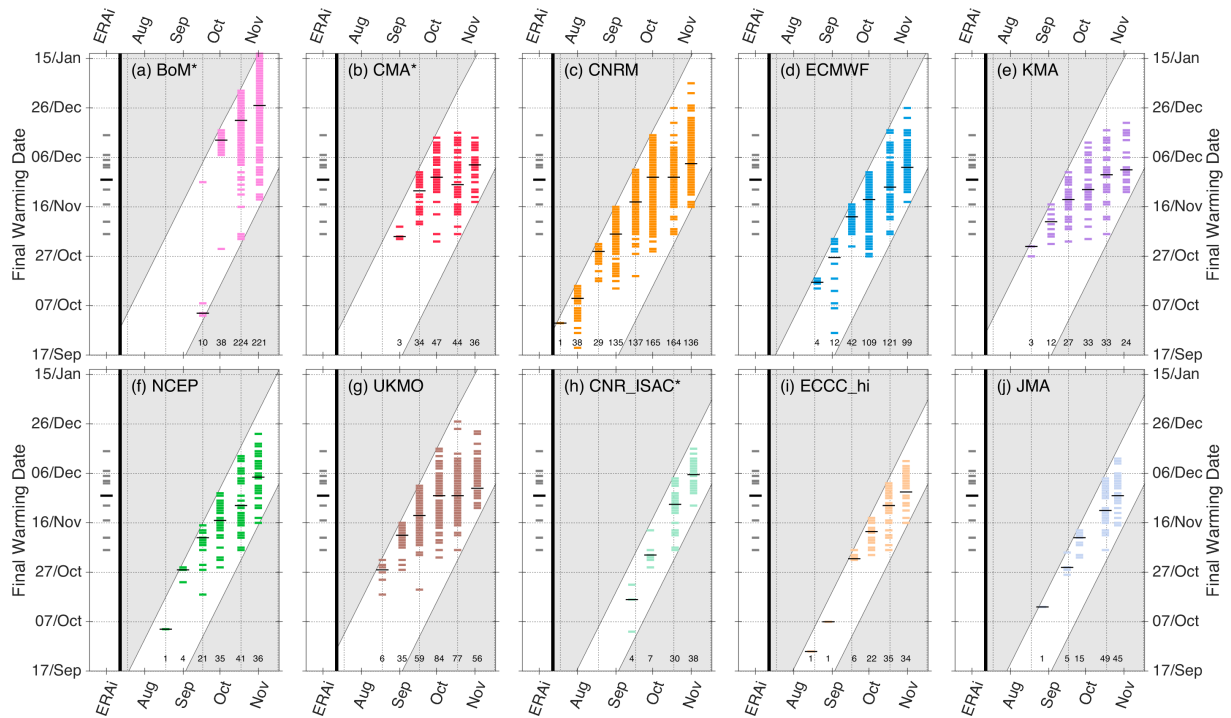
Figure 11. Plots showing the distribution of southern hemisphere final warming dates predicted by each S2S system. The reanalysis final warming dates from ERA-I are shown by the grey strips on the left side of each panel. The systems are composited based on initializations that fall closest to the beginning or middle of each month as specified on the x-axis. The diagonal white region represents the approximate maximum forecast horizon for each system and initialization.

HMCR particularly show nearly flat variations, indicating that they do not simulate an appreciable seasonal cycle in variability. The overestimation of SH variability shown by NCEP from week 4 and beyond in October and November initializations is consistent with this system having the most negative heat flux biases in Figure 9. Overall, because of the limited number of years in the comparison, most of these differences in variability with respect to reanalysis are not significant (not shown).

There are also biases in the timing of the seasonal breakdown of the SH polar vortex. Figure 11 shows histograms of the polar vortex breakdown dates across different initialization dates between August and November. Here we simply define the breakdown date as the first day of easterlies without subsequent return to westerlies at 10 hPa and 60°S. For many of the models, forecasts initialized in early spring produce easterly winds towards the end of the forecast, but at times that are much too early for the breakdown. This is a somewhat surprising result given that such early polar vortex wind reversals are rare in observations and expected to be rare in model simulations (e.g., Jucker et al., 2021). It is unclear, however, whether this behavior represents an early breakdown bias among some of the forecast systems, or whether these events represent SSWs for which the polar vortex would eventually recover if the forecasts continued in time. Regardless, these models produce early events that are generally not consistent with the observational record. Breakdown dates that fall within the ERA-I range are

generally only produced for initializations on or after the first of October, or once the end of the forecasts include the second half of December. An exception is the BoM model, which produces breakdown dates that consistently fall close to the end of its forecasts, resulting in a late breakdown bias for initializations after October. Notably, free-running coupled climate models have a bias towards too-late breakdowns of the SH polar vortex (Butchart et al., 2011; Rao and Garfinkel, 2021b); this suggests that information contained in the October initializations likely help to constrain the S2S models and improve final warming estimates. Although stratospheric ozone is prescribed to climatological values in many S2S forecast systems, the strength of the initialized polar vortex winds in October likely contains information about relevant chemistry-climate feedbacks with stratospheric ozone that are well-correlated to the timing of the breakdown date (Butler and Domeisen, 2021).

4 Discussion and Conclusions

We have performed a comprehensive intercomparison of stratospheric biases in subseasonal forecast systems, with a core focus on systems that contribute to the S2S database (Vitart et al., 2017). Our results show:

– Forecast systems with low-top atmospheric models generally have the largest biases across the diagnostics examined for zonal mean winds and temperatures, the QBO, meridional eddy heat fluxes, and the stratospheric polar vortices.

– Global- and annual-mean warm biases in the stratosphere tend to be most common across the different S2S forecast systems, though this can vary for different regions of the stratosphere in some (e.g., lower versus middle stratosphere).

– Too strong/cold wintertime polar vortices, and too cold extratropical UTLS regions are common features across most of the systems in the zonal mean temperature and zonal wind biases.

– Tropical stratospheric anomalies associated with the QBO tend to decay with lead time to be too weak compared to reanalysis. For high-top systems, this issue is mostly only apparent in the lower stratosphere.

– Stratospheric polar vortex anomalies associated with different phases of the QBO (the "Holton-Tan" relationship) do develop in the forecast systems, but they are generally weaker than in reanalysis.

– In the NH, most S2S forecast systems do not capture the seasonal cycle of extreme vortex event probabilities; for example, the occurrence of SSWs and strong vortex events are underestimated and overestimated, respectively, for weeks 3-4 forecasts initialized in January. Similarly, the S2S systems generally underestimate the magnitude of wind changes associated with observed SSWs and strong vortex events, even at lead times within 2 weeks.

– In the SH, most systems generally underestimate the late-spring variability of the Antarctic polar vortex, particularly for initializations in October and November. However, many systems also simulate reversals in the 10 hPa 60°S zonal mean zonal winds for initializations in August and September, at times of the year when SH final warmings have not occurred in reanalysis.

These biases likely arise due to a combination of factors. While the physical processes that govern the evolution of the stratosphere are relatively well understood, they are generally not fully resolved within atmospheric models, and are instead dependent upon model configurations (e.g., the height of model lids and vertical resolution), and simplified/parameterized processes (e.g., gravity wave drag, and the representation of ozone). The resulting biases can affect both the mean state and variability of the stratosphere, and have potential consequences for subsequent coupling with the troposphere.

Several of the systems with the largest global- and annual-mean warm biases in the stratosphere are those in the "NOAA family", including the high-top NCEP CFSv2, and the low-top GEFSv12 and GFDL-SPEAR; the others include the low-top BoM and CESM2-CAM. The only systems with global- and annual-mean cold biases are the ECMWF and CESM2-WACCM systems. It is unclear whether the warm-biases in the NOAA systems are related to a common cause. While the GEFSv12 and NCEP CFSv2 models use similar physics packages, including ozone physics parameterizations and radiation packages, they do use different dynamical cores (Guan et al., 2021; Saha et al., 2006, 2014); similarly, GEFSv12 and the GFDL-SPEAR share the same FV3-based dynamical core, but SPEAR uses different physics packages and uses prescribed monthly ozone time series (Zhao et al., 2018; Delworth et al., 2020). The fact that the global and annual-mean stratosphere should be in radiative equilibrium poses a strong constraint that biases are likely to be radiative in nature, but model dynamics related to horizontal and vertical resolution can also play a role, particularly at high resolutions (see, e.g., Polichtchouk et al., 2019). We note that the global mean cold biases in the ECMWF system are likely dependent on the specific model cycles, and that more recent versions likely have reduced cold biases following implementation of quintic vertical interpolation in the ECMWF model's semi-Lagrangian numerics (Polichtchouk et al., 2019).

The cold biases in wintertime polar cap temperatures (corresponding to stronger polar stratospheric winds), and cold biases in the extratropical upper troposphere/lower stratosphere are longstanding biases that are similar to what has been documented in other weather and climate models (Charlton-Perez et al., 2013; Bland et al., 2021). The stratospheric polar cap temperature biases generally point to dynamical influences related to planetary wave drag and parameterized gravity wave drag, since wave-mean flow interactions and the ensuing residual circulations are responsible for driving local zonal mean temperatures away from radiative equilibrium. The cold extratropical UTLS biases, on the other hand, are likely to be radiatively driven, related to excessive leakage of water vapor into the lower stratosphere (e.g., Bland et al., 2021). Both of these issues are dependent upon vertical resolution, which likely explains why the biases in the low-top systems (which have fewer levels in the stratosphere and coarser resolution in the UTLS) are generally more severe than those in the high-top systems. Reducing such biases through model improvements is likely to have some impact on forecast skill in both the stratosphere and troposphere, since the latitudinal dependence of the temperature biases affect the distribution of winds (i.e., the tropospheric jets and stratospheric polar vortices). For instance, artificially bias-correcting the extratropical UTLS humidity biases in the ECMWF forecast model was shown to remove the UTLS cold biases, and moderately improve the skill of forecasts over Europe (Hogan et al., 2017).

The gradual decay of QBO anomalies with lead time towards each forecast system's own climatology is consistent with possible issues related to parameterized gravity wave drag. In models, the QBO has been shown to be most sensitive to parameterized non-orographic gravity wave drag (NOGWD) (e.g. Bushell et al., 2020), but other factors such as model vertical diffusion and resolution can also play a role in properly representing and maintaining the QBO, especially in the lower strato-

sphere (Garfinkel et al., 2021; Polichtchouk et al., 2021). However, documenting the specific model configurations and gravity wave parameterizations among the different forecast systems examined herein is beyond the scope of this study. Our results also showed that many of the S2S systems show a "Holton-Tan" response to the QBO wind phase in polar vortex winds and polar cap geopotential heights consistent with observations, but only for the first 2-4 weeks of the forecasts, after which the polar vortex anomalies decay.

To better understand the origin of stratospheric polar vortex biases, we examined the distribution and time evolution of NH lower stratospheric meridional eddy heat fluxes, which are a proxy for the vertical wave activity. These heat fluxes were generally more realistic in the high-top systems; the low-top systems showed considerable negative biases in heat fluxes at week 4 over the Northern Pacific and Scandinavia/Siberia. This suggests that, in combination with their limited representations of the stratosphere, these low-top systems have difficulties simulating realistic Rossby wave activity in the troposphere and/or their propagation and interaction with the mean flow (see, e.g., Schwartz et al., 2021). Thus, other biases documented for the low-top systems (such as those related to polar vortex winds and variability) are likely tied, to some extent, to these deficiencies. We documented similar behavior in the low-top composite of SH eddy heat fluxes, but the biases for the low-top systems were less robust, and were particularly affected by the BoM system having large positive heat flux biases (indicating less upward wave driving in the SH).

Raw forecasts from the S2S systems do not accurately capture the seasonal cycle of boreal SSWs or strong vortex events. Removing the drift in the stratospheric zonal winds through simple bias-correction does improve the probabilities for these events, but primarily only for midwinter occurrences. While perfect prediction of such extreme polar vortex events is not feasible, ideally the statistics for their monthly occurrences derived from the hindcasts would more closely match those from reanalysis, especially at longer lead times and assuming similar levels of tropospheric "noise". Furthermore, the underestimation of the magnitude of wind changes that occur surrounding extreme polar vortex events (even within 1-2 weeks of lead time) suggest that the S2S systems would likely have issues simulating downward coupling associated with such events. The persistence and magnitude of SSWs and strong vortex events is thought to be an important aspect that helps determine whether they lead to coupling with the troposphere (Maycock and Hitchcock, 2015; Karpechko et al., 2017; Charlton-Perez et al., 2018; Domeisen, 2019; White et al., 2019).

In the SH, a large fraction of the S2S systems seem to simulate early breakdowns of the SH polar vortex at 10 hPa, even for forecasts initialized in August and September. Because the forecasts are truncated in time, we cannot say whether these would be considered SSWs or final warmings, but nonetheless our results reveal that false-positive easterly wind events are relatively common in the S2S hindcasts. Such frequent and early reversals in the SH springtime stratospheric circulation is not a phenomenon seen in observations.

Our results show that many of the mean state biases become more sizable with increasing lead time (as expected), especially around weeks 3 to 4. Many of the S2S systems considered herein also make forecasts well beyond 4 weeks; on these extended timescales, such biases are likely to have more substantial impacts on stratosphere-troposphere coupling. While a fully unbiased model will not be possible to achieve, it is still desirable for models to (1) minimize mean state stratospheric biases so that the stratosphere represents a more accurate upper boundary condition for the troposphere, and (2) have similar variability (in

a statistical sense) to observations in the stratosphere so that ensemble spread properly accounts for potential outcomes. For example, forecasts from a model with a strong NH polar vortex bias could simply be post-processed with bias correction to improve the prediction for a SSW (Figure 6); however, if the polar stratospheric winds stay westerly in the actual model simulation, then that would represent a different dynamical regime for stratosphere-troposphere coupling compared to the model actually simulating a transition to stratospheric easterlies (since easterly winds effectively shut off vertical propagation of Rossby waves).

This study primarily focuses on biases among S2S models within the stratosphere. We have shown that large biases are present throughout the stratosphere, linked to a range of stratospheric phenomena in the tropics and the extratropics of both hemispheres. To our knowledge, this is the first systematic assessment of such biases in the stratosphere and of processes affecting the stratosphere in a multi-model study of sub-seasonal to seasonal prediction systems. In a follow-up companion study as part of the same SNAP effort, we more closely examine how biases in the stratosphere, such as those identified herein, are linked to coupling with the troposphere and its predictability.

Data availability. The hindcasts from the S2S database used here are available from https://apps.ecmwf.int/datasets/data/s2s/ under the "Reforecasts" S2S set. The NOAA GEFSv12 hindcasts can be obtained from https://registry.opendata.aws/noaa-gefs-reforecast/. Hindcasts for CESM2-CAM are available at https://www.earthsystemgrid.org/dataset/ucar.cgd.cesm2.s2s_hindcasts.html, while those for CESM2-WACCM are from https://www.earthsystemgrid.org/dataset/ucar.cgd.cesm2-waccm.s2s_hindcasts.html. Data for the GFDL-SPEAR can be made available upon request.

Author contributions. ZDL organized and led the SNAP effort leading to this paper, with help from CIG and AHB. ZDL, AHB, CIG, and DIVD drafted the manuscript and made revisions with feedback from other coauthors. CJW compiled all the figures in the paper and supplement with data provided from the following collaborators/analyses: HK, and S-WS performed the analyses for Figures 1 and 2. CIG and CS performed the analyses for Figures 3 and 4. CIG, CS, FMP, JG-S, AdlC, and NC performed the analyses for Figures 5 and 9. SHL performed the analysis for Figure 6. RW-YW and DIVD performed the analysis for Figure 7, including with data for the GEFS and CESM2 systems provided by ZDL. NPH and CJW performed the analysis for Figure 8 with code provided by ZDL. MO, MJ, and IP performed the analyses for Figures 10 and 11. These analyses were all done for the systems provided in the S2S database; ZDL performed the same analyses for Figures 1, 2, 3, and 6 for the GEFSv12 and CESM2 systems. PL and LJ performed the same analyses for Figures 1 and 2 for the GFDL-SPEAR system. All the listed coauthors were active participants in this SNAP community effort and provided comments on the draft and revised manuscript.

Competing interests. Daniela Domeisen is a member of the editorial board of Weather and Climate Dynamics.

Acknowledgements. This work uses S2S Project data. S2S is a joint initiative of the World Weather Research Programme (WWRP) and the World Climate Research Programme (WCRP). This work was initiated by the Stratospheric Network for the Assessment of Predictability (SNAP), a joint activity of SPARC (WCRP) and the S2S Project (WWRP/WCRP).

The work of R.W. is funded through ETH grant ETH-05 19-1. Support from the Swiss National Science Foundation through projects PP00P2_170523 and PP00P2_198896 to D.D. is gratefully acknowledged. C.I.G. and C.S. are supported by the ISF-NSFC joint research program (grant No.3259/19). The work of M.O. was supported by UBACyT20020170100428BA and PICT-2018-03046 projects. The work of A.dlC. is funded by the Spanish Ministry of Science through project PID2019-109107GB-I00. B.A and N.C. acknowledge the support of the Spanish Ministry of Science and Innovation through the JeDiS (RTI2018-096402-B-I00) project. FMP and JG-S have been partially supported by the Spanish ATLANTE project (PID2019-110234RB-C21) and "Ramón y Cajal" programme (RYC-2016-21181), respectively. N.P.H. and C.J.W. are supported by UK Natural Environment Research Council (NERC) grant number NE/S00985X/1. C.J.W. is also supported by a Royal Society University Research Fellowship UF160545. S.-W.S. and H.K. are supported by the Basic Science Research Program through the National Research Foundation of Korea (2017R1E1A1A01074889). This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research (BER), Regional and Global Model Analysis (RGMA) component of the Earth and Environmental System Modeling Program under Award Number DE-SC0022070 and National Science Foundation (NSF) IA 1947282. This work was also supported by the National Center for Atmospheric Research (NCAR), which is a major facility sponsored by the NSF under Cooperative Agreement No. 1852977. P.L. is supported by award NA18OAR4320123 from the National Oceanic and Atmospheric Administration (NOAA), U.S. Department of Commerce. ZDL was partially supported under NOAA Award NA20NWS4680051; ZDL and JP also acknowledge support from US Federally Appropriated Funds. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA, or the U.S. Department of Commerce.

References

640 Afargan-Gerstman, H. and Domeisen, D. I. V.: Pacific Modulation of the North Atlantic Storm Track Response to Sudden Stratospheric Warming Events, Geophysical Research Letters, 47, e2019GL085 007, https://doi.org/10.1029/2019GL085007, 2020.

Ambaum, M. H. P. and Hoskins, B. J.: The NAO Troposphere–Stratosphere Connection, Journal of Climate, 15, 1969–1978, https://doi.org/10.1175/1520-0442(2002)015<1969:TNTSC>2.0.CO;2, 2002.

Andrews, D. G., Leovy, C. B., and Holton, J. R.: Middle Atmosphere Dynamics, Academic Press, 1987.

645 Anstey, J. A. and Shepherd, T. G.: High-Latitude Influence of the Quasi-Biennial Oscillation, Quarterly Journal of the Royal Meteorological Society, 140, 1–21, https://doi.org/10.1002/qj.2132, 2014.

Anstey, J. A., Simpson, I. R., Richter, J. H., Naoe, H., Taguchi, M., Serva, F., Gray, L. J., Butchart, N., Hamilton, K., Osprey, S., Bellprat, O., Braesicke, P., Bushell, A. C., Cagnazzo, C., Chen, C.-C., Chun, H.-Y., Garcia, R. R., Holt, L., Kawatani, Y., Kerzenmacher, T., Kim, Y.-H., Lott, F., McLandress, C., Scinocca, J., Stockdale, T. N., Versick, S., Watanabe, S., Yoshida, K., and Yukimoto, S.: Teleconnections 650 of the Quasi-Biennial Oscillation in a Multi-Model Ensemble of QBO-resolving Models, Quarterly Journal of the Royal Meteorological Society, https://doi.org/10.1002/qj.4048, 2021.

Ayarzagüena, B., Charlton-Perez, A. J., Butler, A. H., Hitchcock, P., Simpson, I. R., Polvani, L. M., Butchart, N., Gerber, E. P., Gray, L., Hassler, B., Lin, P., Lott, F., Manzini, E., Mizuta, R., Orbe, C., Osprey, S., Saint-Martin, D., Sigmond, M., Taguchi, M., Volodin, E. M., and Watanabe, S.: Uncertainty in the Response of Sudden Stratospheric Warmings and Stratosphere-Troposphere Cou-655 pling to Quadrupled CO2 Concentrations in CMIP6 Models, Journal of Geophysical Research: Atmospheres, 125, e2019JD032 345, https://doi.org/10.1029/2019JD032345, 2020.

Baldwin, M. P. and Dunkerton, T. J.: Stratospheric Harbingers of Anomalous Weather Regimes, Science, 294, 581–584, https://doi.org/10.1126/science.1063315, 2001.

Baldwin, M. P., Gray, L. J., Dunkerton, T. J., Hamilton, K., Haynes, P. H., Randel, W. J., Holton, J. R., Alexander, M. J., Hirota, I., Horinouchi, 660 T., Jones, D. B. A., Kinnersley, J. S., Marquardt, C., Sato, K., and Takahashi, M.: The Quasi-Biennial Oscillation, Reviews of Geophysics, 39, 179–229, https://doi.org/10.1029/1999RG000073, 2001.

Baldwin, M. P., Ayarzagüena, B., Birner, T., Butchart, N., Butler, A. H., Charlton-Perez, A. J., Domeisen, D. I. V., Garfinkel, C. I., Garny, H., Gerber, E. P., Hegglin, M. I., Langematz, U., and Pedatella, N. M.: Sudden Stratospheric Warmings, Reviews of Geophysics, 59, e2020RG000 708, https://doi.org/10.1029/2020RG000708, 2021.

665 Black, R. X., McDaniel, B. A., and Robinson, W. A.: Stratosphere–Troposphere Coupling during Spring Onset, Journal of Climate, 19, 4891–4901, https://doi.org/10.1175/JCLI3907.1, 2006.

Bland, J., Gray, S., Methven, J., and Forbes, R.: Characterising Extratropical Near-Tropopause Analysis Humidity Biases and Their Radiative Effects on Temperature Forecasts, Quarterly Journal of the Royal Meteorological Society, https://doi.org/10.1002/qj.4150, 2021.

Bushell, A. C., Anstey, J. A., Butchart, N., Kawatani, Y., Osprey, S. M., Richter, J. H., Serva, F., Braesicke, P., Cagnazzo, C., Chen, C.-C., 670 Chun, H.-Y., Garcia, R. R., Gray, L. J., Hamilton, K., Kerzenmacher, T., Kim, Y.-H., Lott, F., McLandress, C., Naoe, H., Scinocca, J., Smith, A. K., Stockdale, T. N., Versick, S., Watanabe, S., Yoshida, K., and Yukimoto, S.: Evaluation of the Quasi-Biennial Oscillation in Global Climate Models for the SPARC QBO-initiative, Quarterly Journal of the Royal Meteorological Society, pp. 1–31, https://doi.org/10.1002/qj.3765, 2020.

Butchart, N., Charlton-Perez, A. J., Cionni, I., Hardiman, S. C., Haynes, P. H., Krüger, K., Kushner, P. J., Newman, P. A., Osprey, S. M., 675 Perlwitz, J., Sigmond, M., Wang, L., Akiyoshi, H., Austin, J., Bekki, S., Baumgaertner, A., Braesicke, P., Brühl, C., Chipperfield, M.,

Dameris, M., Dhomse, S., Eyring, V., Garcia, R., Garny, H., Jöckel, P., Lamarque, J.-F., Marchand, M., Michou, M., Morgenstern, O., Nakamura, T., Pawson, S., Plummer, D., Pyle, J., Rozanov, E., Scinocca, J., Shepherd, T. G., Shibata, K., Smale, D., Teyssèdre, H., Tian, W., Waugh, D., and Yamashita, Y.: Multimodel Climate and Variability of the Stratosphere, Journal of Geophysical Research: Atmospheres, 116, https://doi.org/10.1029/2010JD014995, 2011.

680 Butler, A., Charlton-Perez, A., Domeisen, D. I. V., Garfinkel, C., Gerber, E. P., Hitchcock, P., Karpechko, A. Y., Maycock, A. C., Sigmond, M., Simpson, I., and Son, S.-W.: Chapter 11 - Sub-seasonal Predictability and the Stratosphere, in: Sub-Seasonal to Seasonal Prediction, edited by Robertson, A. W. and Vitart, F., pp. 223–241, Elsevier, https://doi.org/10.1016/B978-0-12-811714-9.00011-5, 2019a.

Butler, A. H. and Domeisen, D. I. V.: The Wave Geometry of Final Stratospheric Warming Events, Weather and Climate Dynamics, 2, 453–474, https://doi.org/10.5194/wcd-2-453-2021, 2021.

685 Butler, A. H., Arribas, A., Athanassiadou, M., Baehr, J., Calvo, N., Charlton-Perez, A., Déqué, M., Domeisen, D. I. V., Fröhlich, K., Hendon, H., Imada, Y., Ishii, M., Iza, M., Karpechko, A. Y., Kumar, A., MacLachlan, C., Merryfield, W. J., Müller, W. A., O'Neill, A., Scaife, A. A., Scinocca, J., Sigmond, M., Stockdale, T. N., and Yasuda, T.: The Climate-system Historical Forecast Project: Do Stratosphere-Resolving Models Make Better Seasonal Climate Predictions in Boreal Winter?, Quarterly Journal of the Royal Meteorological Society, 142, 1413–1427, https://doi.org/10.1002/qj.2743, 2016.

690 Butler, A. H., Sjoberg, J. P., Seidel, D. J., and Rosenlof, K. H.: A Sudden Stratospheric Warming Compendium, Earth System Science Data, 9, 63–76, https://doi.org/10.5194/essd-9-63-2017, 2017.

Butler, A. H., Charlton-Perez, A., Domeisen, D. I., Simpson, I. R., and Sjoberg, J.: Predictability of Northern Hemisphere Final Stratospheric Warmings and Their Surface Impacts, Geophysical Research Letters, 46, 10 578–10 588, https://doi.org/10.1029/2019GL083346, 2019b.

Byrne, N. J., Shepherd, T. G., Woollings, T., and Plumb, R. A.: Nonstationarity in Southern Hemisphere Climate Variability Associated with
695 the Seasonal Breakdown of the Stratospheric Polar Vortex, Journal of Climate, 30, 7125–7139, https://doi.org/10.1175/JCLI-D-17-0097.1, 2017.

Byrne, N. J., Shepherd, T. G., and Polichtchouk, I.: Subseasonal-to-Seasonal Predictability of the Southern Hemisphere Eddy-Driven Jet During Austral Spring and Early Summer, Journal of Geophysical Research: Atmospheres, 124, 6841–6855, https://doi.org/10.1029/2018JD030173, 2019.

700 Calvo, N., Giorgetta, M. A., and Peña-Ortiz, C.: Sensitivity of the Boreal Winter Circulation in the Middle Atmosphere to the Quasi-Biennial Oscillation in MAECHAM5 Simulations, Journal of Geophysical Research: Atmospheres, 112, https://doi.org/10.1029/2006JD007844, 2007.

Charlton, A. J. and Polvani, L. M.: A New Look at Stratospheric Sudden Warmings. Part I: Climatology and Modeling Benchmarks, Journal of Climate, 20, 449–469, https://doi.org/10.1175/JCLI3996.1, 2007.

705 Charlton-Perez, A. J., Baldwin, M. P., Birner, T., Black, R. X., Butler, A. H., Calvo, N., Davis, N. A., Gerber, E. P., Gillett, N., Hardiman, S., Kim, J., Krüger, K., Lee, Y.-Y., Manzini, E., McDaniel, B. A., Polvani, L., Reichler, T., Shaw, T. A., Sigmond, M., Son, S.-W., Toohey, M., Wilcox, L., Yoden, S., Christiansen, B., Lott, F., Shindell, D., Yukimoto, S., and Watanabe, S.: On the Lack of Stratospheric Dynamical Variability in Low-Top Versions of the CMIP5 Models, Journal of Geophysical Research: Atmospheres, 118, 2494–2505, https://doi.org/10.1002/jgrd.50125, 2013.

710 Charlton-Perez, A. J., Ferranti, L., and Lee, R. W.: The Influence of the Stratospheric State on North Atlantic Weather Regimes, Quarterly Journal of the Royal Meteorological Society, 144, 1140–1151, https://doi.org/10.1002/qj.3280, 2018.

Charlton-Perez, A. J., Ferranti, L., and Lee, R. W.: The Influence of the Stratospheric State on North Atlantic Weather Regimes, Quarterly Journal of the Royal Meteorological Society, 144, 1140–1151, https://doi.org/10.1002/qj.3280, 2018.

Cohen, J., Agel, L., Barlow, M., Garfinkel, C. I., and White, I.: Linking Arctic Variability and Change with Extreme Winter Weather in the United States, Science, 373, 1116–1121, https://doi.org/10.1126/science.abi9167, 2021.

Dai, Y. and Hitchcock, P.: Understanding the Basin Asymmetry in Surface Response to Sudden Stratospheric Warmings from an Ocean–Atmosphere Coupled Perspective, Journal of Climate, 34, 8683–8698, https://doi.org/10.1175/JCLI-D-21-0314.1, 2021.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim Reanalysis: Configuration and Performance of the Data Assimilation System, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, https://doi.org/10.1002/qj.828, 2011.

Delworth, T. L., Cooke, W. F., Adcroft, A., Bushuk, M., Chen, J.-H., Dunne, K. A., Ginoux, P., Gudgel, R., Hallberg, R. W., Harris, L., Harrison, M. J., Johnson, N., Kapnick, S. B., Lin, S.-J., Lu, F., Malyshev, S., Milly, P. C., Murakami, H., Naik, V., Pascale, S., Paynter, D., Rosati, A., Schwarzkopf, M., Shevliakova, E., Underwood, S., Wittenberg, A. T., Xiang, B., Yang, X., Zeng, F., Zhang, H., Zhang, L., and Zhao, M.: SPEAR: The Next Generation GFDL Modeling System for Seasonal to Multidecadal Prediction and Projection, Journal of Advances in Modeling Earth Systems, 12, e2019MS001 895, https://doi.org/10.1029/2019MS001895, 2020.

Domeisen, D. I.: Estimating the Frequency of Sudden Stratospheric Warming Events From Surface Observations of the North Atlantic Oscillation, Journal of Geophysical Research: Atmospheres, 124, 3180–3194, https://doi.org/10.1029/2018JD030077, 2019.

Domeisen, D. I., Butler, A. H., Charlton-Perez, A. J., Ayarzagüena, B., Baldwin, M. P., Dunn-Sigouin, E., Furtado, J. C., Garfinkel, C. I., Hitchcock, P., Karpechko, A. Y., Kim, H., Knight, J., Lang, A. L., Lim, E.-P., Marshall, A., Roff, G., Schwartz, C., Simpson, I. R., Son, S.-W., and Taguchi, M.: The Role of the Stratosphere in Subseasonal to Seasonal Prediction: 1. Predictability of the Stratosphere, Journal of Geophysical Research: Atmospheres, 125, e2019JD030 920, https://doi.org/10.1029/2019JD030920, 2020a.

Domeisen, D. I. V. and Butler, A. H.: Stratospheric Drivers of Extreme Events at the Earth's Surface, Commun Earth Environ, 1, 1–8, https://doi.org/10.1038/s43247-020-00060-z, 2020.

Domeisen, D. I. V., Butler, A. H., Fröhlich, K., Bittner, M., Müller, W. A., and Baehr, J.: Seasonal Predictability over Europe Arising from El Niño and Stratospheric Variability in the MPI-ESM Seasonal Prediction System, Journal of Climate, 28, 256–271, https://doi.org/10.1175/JCLI-D-14-00207.1, 2015.

Domeisen, D. I. V., Butler, A. H., Charlton-Perez, A. J., Ayarzagüena, B., Baldwin, M. P., Dunn-Sigouin, E., Furtado, J. C., Garfinkel, C. I., Hitchcock, P., Karpechko, A. Y., Kim, H., Knight, J., Lang, A. L., Lim, E.-P., Marshall, A., Roff, G., Schwartz, C., Simpson, I. R., Son, S.-W., and Taguchi, M.: The Role of the Stratosphere in Subseasonal to Seasonal Prediction: 2. Predictability Arising From Stratosphere-Troposphere Coupling, Journal of Geophysical Research: Atmospheres, 125, e2019JD030 923, https://doi.org/10.1029/2019JD030923, 2020b.

Domeisen, D. I. V., Grams, C. M., and Papritz, L.: The Role of North Atlantic–European Weather Regimes in the Surface Impact of Sudden Stratospheric Warming Events, Weather and Climate Dynamics, 1, 373–388, https://doi.org/10.5194/wcd-1-373-2020, 2020c.

Elsbury, D., Peings, Y., and Magnusdottir, G.: Variation in the Holton–Tan Effect by Longitude, Quarterly Journal of the Royal Meteorological Society, 147, 1767–1787, https://doi.org/10.1002/qj.3993, 2021.

Fujiwara, M., Manney, G. L., Gray, L. J., and Wright, J. S.: SPARC Reanalysis Intercomparison Project (S-RIP) Final Report, SPARC Report No. 10, https://doi.org/10.17874/800dee57d13, 2021.

Garfinkel, C. I., Hartmann, D. L., and Sassi, F.: Tropospheric Precursors of Anomalous Northern Hemisphere Stratospheric Polar Vortices, Journal of Climate, 23, 3282–3299, https://doi.org/10.1175/2010JCLI3010.1, 2010.

Garfinkel, C. I., Shaw, T. A., Hartmann, D. L., and Waugh, D. W.: Does the Holton–Tan Mechanism Explain How the Quasi-Biennial Oscillation Modulates the Arctic Polar Vortex?, Journal of the Atmospheric Sciences, 69, 1713–1733, https://doi.org/10.1175/JAS-D-11-0209.1, 2012.

Garfinkel, C. I., Schwartz, C., Domeisen, D. I. V., Son, S.-W., Butler, A. H., and White, I. P.: Extratropical Atmospheric Predictability From the Quasi-Biennial Oscillation in Subseasonal Forecast Models, Journal of Geophysical Research: Atmospheres, 123, 7855–7866, https://doi.org/10.1029/2018JD028724, 2018.

Garfinkel, C. I., Gerber, E. P., Shamir, O., Rao, J., Jucker, M., White, I., and Paldor, N.: A QBO Cookbook: Sensitivity of the Quasi-Biennial Oscillation to Resolution, Resolved Waves, and Parameterized Gravity Waves, Journal of Advances in Modeling Earth Systems, n/a, e2021MS002 568, https://doi.org/10.1029/2021MS002568, 2021.

Geller, M. A., Zhou, T., Shindell, D., Ruedy, R., Aleinov, I., Nazarenko, L., Tausnev, N. L., Kelley, M., Sun, S., Cheng, Y., Field, R. D., and Faluvegi, G.: Modeling the QBO—Improvements Resulting from Higher-Model Vertical Resolution, Journal of Advances in Modeling Earth Systems, 8, 1092–1105, https://doi.org/10.1002/2016MS000699, 2016.

Gerber, E. P. and Martineau, P.: Quantifying the Variability of the Annular Modes: Reanalysis Uncertainty vs. Sampling Uncertainty, Atmospheric Chemistry and Physics, 18, 17 099–17 117, https://doi.org/10.5194/acp-18-17099-2018, 2018.

Gerber, E. P., Butler, A., Calvo, N., Charlton-Perez, A., Giorgetta, M., Manzini, E., Perlwitz, J., Polvani, L. M., Sassi, F., Scaife, A. A., Shaw, T. A., Son, S.-W., and Watanabe, S.: Assessing and Understanding the Impact of Stratospheric Dynamics and Variability on the Earth System, Bulletin of the American Meteorological Society, 93, 845–859, https://doi.org/10.1175/BAMS-D-11-00145.1, 2012.

Gettelman, A., Mills, M. J., Kinnison, D. E., Garcia, R. R., Smith, A. K., Marsh, D. R., Tilmes, S., Vitt, F., Bardeen, C. G., McInerny, J., Liu, H.-L., Solomon, S. C., Polvani, L. M., Emmons, L. K., Lamarque, J.-F., Richter, J. H., Glanville, A. S., Bacmeister, J. T., Phillips, A. S., Neale, R. B., Simpson, I. R., DuVivier, A. K., Hodzic, A., and Randel, W. J.: The Whole Atmosphere Community Climate Model Version 6 (WACCM6), Journal of Geophysical Research: Atmospheres, 124, 12 380–12 403, https://doi.org/10.1029/2019JD030943, 2019.

Gray, L. J., Anstey, J. A., Kawatani, Y., Lu, H., Osprey, S., and Schenzinger, V.: Surface Impacts of the Quasi Biennial Oscillation, Atmospheric Chemistry and Physics, 18, 8227–8247, https://doi.org/10.5194/acp-18-8227-2018, 2018.

Guan, H., Zhu, Y., Sinsky, E., Fu, B., Li, W., Zhou, X., Xue, X., Hou, D., Peng, J., Nageswararao, M. M., Tallapgragada, V., Hamill, T. M., Whitaker, J. S., Pegion, P., Frederick, S., Rosencrans, M., and Kumar, A.: GEFSv12 Reforecast Dataset for Supporting Subseasonal and Hydrometeorological Applications, submitted to Monthly Weather Review, 2021.

Hamill, T. M., Whitaker, J. S., Shlyaeva, A., Bates, G., Fredrick, S., Pegion, P., Sinsky, E., Zhu, Y., Tallapragada, V., Guan, H., Zhou, X., and Woollen, J.: The Reanalysis for the Global Ensemble Forecast System, Version 12., Monthly Weather Review, -1, https://doi.org/10.1175/MWR-D-21-0023.1, 2021.

Haynes, P., Hitchcock, P., Hitchman, M., Yoden, S., Hendon, H., Kiladis, G., Kodera, K., and Simpson, I.: The Influence of the Stratosphere on the Tropical Troposphere, Journal of the Meteorological Society of Japan. Ser. II, 99, 803–845, https://doi.org/10.2151/jmsj.2021-040, 2021.

Hio, Y. and Yoden, S.: Interannual Variations of the Seasonal March in the Southern Hemisphere Stratosphere for 1979–2002 and Characterization of the Unprecedented Year 2002, Journal of the Atmospheric Sciences, 62, 567–580, https://doi.org/10.1175/JAS-3333.1, 2005.

Hogan, R., Ahlgrimm, M., Balsamo, G., Beljaars, A., Berrisford, P., Bozzo, A., Di Giuseppe, F., Forbes, R. M., Haiden, T., Lang, S., Mayer, M., Polichtchouk, I., Sandu, I., Vitart, F., and Wedi, N.: Radiation in Numerical Weather Prediction, ECMWF Technical Memoranda, https://doi.org/10.21957/2bd5dkj8x, 2017.

Holton, J. R. and Tan, H.-C.: The Influence of the Equatorial Quasi-Biennial Oscillation on the Global Circulation at 50 Mb, Journal of the Atmospheric Sciences, 37, 2200–2208, https://doi.org/10.1175/1520-0469(1980)037<2200:TIOTEQ>2.0.CO;2, 1980.

Jiménez-Esteve, B. and Domeisen, D. I. V.: The Tropospheric Pathway of the ENSO–North Atlantic Teleconnection, Journal of Climate, 31, 4563–4584, https://doi.org/10.1175/JCLI-D-17-0716.1, 2018.

Jucker, M., Reichler, T., and Waugh, D. W.: How Frequent Are Antarctic Sudden Stratospheric Warmings in Present and Future Climate?, Geophysical Research Letters, 48, e2021GL093 215, https://doi.org/10.1029/2021GL093215, 2021.

Karpechko, A. Y., Hitchcock, P., Peters, D. H. W., and Schneidereit, A.: Predictability of Downward Propagation of Major Sudden Stratospheric Warmings, Quarterly Journal of the Royal Meteorological Society, 143, 1459–1470, https://doi.org/10.1002/qj.3017, 2017.

Karpechko, A. Y., Tyrrell, N. L., and Rast, S.: Sensitivity of QBO Teleconnection to Model Circulation Biases, Quarterly Journal of the Royal Meteorological Society, 147, 2147–2159, https://doi.org/10.1002/qj.4014, 2021.

Kim, H., Richter, J. H., and Martin, Z.: Insignificant QBO-MJO Prediction Skill Relationship in the SubX and S2S Subseasonal Reforecasts, Journal of Geophysical Research: Atmospheres, 124, 12 655–12 666, https://doi.org/10.1029/2019JD031416, 2019.

Kim, H., Son, S.-W., and Yoo, C.: QBO Modulation of the MJO-Related Precipitation in East Asia, Journal of Geophysical Research: Atmospheres, 125, e2019JD031 929, https://doi.org/10.1029/2019JD031929, 2020.

Kolstad, E. W., Wulff, C. O., Domeisen, D. I. V., and Woollings, T.: Tracing North Atlantic Oscillation Forecast Errors to Stratospheric Origins, Journal of Climate, 33, 9145–9157, https://doi.org/10.1175/JCLI-D-20-0270.1, 2020.

Lee, J. C. K. and Klingaman, N. P.: The Effect of the Quasi-Biennial Oscillation on the Madden–Julian Oscillation in the Met Office Unified Model Global Ocean Mixed Layer Configuration, Atmospheric Science Letters, 19, e816, https://doi.org/10.1002/asl.816, 2018.

Lim, E.-P., Hendon, H. H., and Thompson, D. W. J.: Seasonal Evolution of Stratosphere-Troposphere Coupling in the Southern Hemisphere and Implications for the Predictability of Surface Climate, Journal of Geophysical Research: Atmospheres, 123, 12,002–12,016, https://doi.org/10.1029/2018JD029321, 2018.

Lim, Y., Son, S.-W., Marshall, A. G., Hendon, H. H., and Seo, K.-H.: Influence of the QBO on MJO Prediction Skill in the Subseasonal-to-Seasonal Prediction Models, Clim Dyn, 53, 1681–1695, https://doi.org/10.1007/s00382-019-04719-y, 2019.

Limpasuvan, V., Hartmann, D. L., Thompson, D. W. J., Jeev, K., and Yung, Y. L.: Stratosphere-Troposphere Evolution during Polar Vortex Intensification, Journal of Geophysical Research: Atmospheres, 110, https://doi.org/10.1029/2005JD006302, 2005.

Long, C. S., Fujiwara, M., Davis, S., Mitchell, D. M., and Wright, C. J.: Climatology and Interannual Variability of Dynamic Variables in Multiple Reanalyses Evaluated by the SPARC Reanalysis Intercomparison Project (S-RIP), Atmospheric Chemistry and Physics, 17, 14 593–14 629, https://doi.org/10.5194/acp-17-14593-2017, 2017.

Marshall, A. G. and Scaife, A. A.: Impact of the QBO on Surface Winter Climate, Journal of Geophysical Research: Atmospheres, 114, https://doi.org/10.1029/2009JD011737, 2009.

Martin, Z., Orbe, C., Wang, S., and Sobel, A.: The MJO–QBO Relationship in a GCM with Stratospheric Nudging, Journal of Climate, 34, 4603–4624, https://doi.org/10.1175/JCLI-D-20-0636.1, 2021a.

Martin, Z., Son, S.-W., Butler, A., Hendon, H., Kim, H., Sobel, A., Yoden, S., and Zhang, C.: The Influence of the Quasi-Biennial Oscillation on the Madden–Julian Oscillation, Nat Rev Earth Environ, 2, 477–489, https://doi.org/10.1038/s43017-021-00173-9, 2021b.

825 Matthias, V. and Kretschmer, M.: The Influence of Stratospheric Wave Reflection on North American Cold Spells, Monthly Weather Review, 148, 1675–1690, https://doi.org/10.1175/MWR-D-19-0339.1, 2020.

Maycock, A. C. and Hitchcock, P.: Do Split and Displacement Sudden Stratospheric Warmings Have Different Annular Mode Signatures?, Geophysical Research Letters, 42, 10,943–10,951, https://doi.org/10.1002/2015GL066754, 2015.

McLandress, C., Shepherd, T. G., Polavarapu, S., and Beagley, S. R.: Is Missing Orographic Gravity Wave Drag near 60°S the
830 Cause of the Stratospheric Zonal Wind Biases in Chemistry–Climate Models?, Journal of the Atmospheric Sciences, 69, 802–818, https://doi.org/10.1175/JAS-D-11-0159.1, 2012.

Mechoso, C. R., Hartmann, D. L., and Farrara, J. D.: Climatology and Interannual Variability of Wave, Mean-Flow Interaction in the Southern Hemisphere, Journal of the Atmospheric Sciences, 42, 2189–2206, https://doi.org/10.1175/1520-0469(1985)042<2189:CAIVOW>2.0.CO;2, 1985.

835 Merryfield, W. J., Baehr, J., Batté, L., Becker, E. J., Butler, A. H., Coelho, C. A. S., Danabasoglu, G., Dirmeyer, P. A., Doblas-Reyes, F. J., Domeisen, D. I. V., Ferranti, L., Ilynia, T., Kumar, A., Müller, W. A., Rixen, M., Robertson, A. W., Smith, D. M., Takaya, Y., Tuma, M., Vitart, F., White, C. J., Alvarez, M. S., Ardilouze, C., Attard, H., Baggett, C., Balmaseda, M. A., Beraki, A. F., Bhattacharjee, P. S., Bilbao, R., de Andrade, F. M., DeFlorio, M. J., Díaz, L. B., Ehsan, M. A., Fragkoulidis, G., Grainger, S., Green, B. W., Hell, M. C., Infanti, J. M., Isensee, K., Kataoka, T., Kirtman, B. P., Klingaman, N. P., Lee, J.-Y., Mayer, K., McKay, R., Mecking, J. V., Miller,
840 D. E., Neddermann, N., Ng, C. H. J., Ossó, A., Pankatz, K., Peatman, S., Pegion, K., Perlwitz, J., Recalde-Coronel, G. C., Reintges, A., Renkl, C., Solaraju-Murali, B., Spring, A., Stan, C., Sun, Y. Q., Tozer, C. R., Vigaud, N., Woolnough, S., and Yeager, S.: Current and Emerging Developments in Subseasonal to Decadal Prediction, Bulletin of the American Meteorological Society, 101, E869–E896, https://doi.org/10.1175/BAMS-D-19-0037.1, 2020.

Messori, G., Kretschmer, M., Lee, S. H., and Matthias, V.: Stratospheric Wave Reflection Events Modulate North American Weather Regimes
845 and Cold Spells, Weather and Climate Dynamics Discussions, pp. 1–29, https://doi.org/10.5194/wcd-2022-18, 2022.

Oh, J., Son, S.-W., Choi, J., Lim, E.-P., Garfinkel, C., Hendon, H., Kim, Y., and Kang, H.-S.: Impact of Stratospheric Ozone on the Subseasonal Prediction in the Southern Hemisphere Spring, Progress in Earth and Planetary Science, 9, 25, https://doi.org/10.1186/s40645-022-00485-4, 2022.

Olaguer, E. P., Yang, H., and Tung, K. K.: A Reexamination of the Radiative Balance of the Stratosphere, Journal of the Atmospheric
850 Sciences, 49, 1242–1263, https://doi.org/10.1175/1520-0469(1992)049<1242:AROTRB>2.0.CO;2, 1992.

Park, C.-H., Son, S.-W., Lim, Y., and Choi, J.: Quasi-Biennial Oscillation-Related Surface Air Temperature Change over the Western North Pacific in Late Winter, International Journal of Climatology, n/a, https://doi.org/10.1002/joc.7470, 2022.

Pegion, K., Kirtman, B. P., Becker, E., Collins, D. C., LaJoie, E., Burgman, R., Bell, R., DelSole, T., Min, D., Zhu, Y., Li, W., Sinsky, E., Guan, H., Gottschalck, J., Metzger, E. J., Barton, N. P., Achuthavarier, D., Marshak, J., Koster, R. D., Lin, H., Gagnon, N., Bell,
855 M., Tippett, M. K., Robertson, A. W., Sun, S., Benjamin, S. G., Green, B. W., Bleck, R., and Kim, H.: The Subseasonal Experiment (SubX): A Multimodel Subseasonal Prediction Experiment, Bulletin of the American Meteorological Society, 100, 2043–2060, https://doi.org/10.1175/BAMS-D-18-0270.1, 2019.

Polichtchouk, I., Stockdale, T., Bechtold, P., Diamantakis, M., Malardel, S., Sandu, I., Vána, F., and Wedi, N.: Control on Stratospheric Temperature in IFS: Resolution and Vertical Advection, ECMWF Technical Memoranda, https://doi.org/10.21957/cz3t12t7e, 2019.

860 Polichtchouk, I., Diamantakis, M., and Vána, F.: Quintic Vertical Interpolation Improves Forecasts of the Stratosphere, ECMWF Newsletter, pp. 23–26, https://doi.org/10.21957/ip52yu935j, 2020.

Polichtchouk, I., Bechtold, P., Bonavita, M., Forbes, R., Healy, S., Hogan, R., Laloyaux, P., Rennie, M., Stockdale, T., Wedi, N., Diamantakis, M., Flemming, J., English, S., Isaksen, L., Vána, F., Gisinger, S., and Byrne, N.: Stratospheric Modelling and Assimilation, ECMWF Technical Memoranda, 877, https://doi.org/10.21957/25hegfoq, 2021.

865 Polvani, L. M. and Waugh, D. W.: Upward Wave Activity Flux as a Precursor to Extreme Stratospheric Events and Subsequent Anomalous Surface Weather Regimes, Journal of Climate, 17, 3548–3554, https://doi.org/10.1175/1520-0442(2004)017<3548:UWAFAA>2.0.CO;2, 2004.

Portal, A., Ruggieri, P., Palmeiro, F. M., García-Serrano, J., Domeisen, D. I. V., and Gualdi, S.: Seasonal Prediction of the Boreal Winter Stratosphere, Clim Dyn, https://doi.org/10.1007/s00382-021-05787-9, 2021.

870 Randel, W. J.: The Seasonal Evolution of Planetary Waves in the Southern Hemisphere Stratosphere and Troposphere, Quarterly Journal of the Royal Meteorological Society, 114, 1385–1409, https://doi.org/10.1002/qj.49711448403, 1988.

Rao, J. and Garfinkel, C. I.: CMIP5/6 Models Project Little Change in the Statistical Characteristics of Sudden Stratospheric Warmings in the 21st Century, Environ. Res. Lett., 16, 034 024, https://doi.org/10.1088/1748-9326/abd4fe, 2021a.

Rao, J. and Garfinkel, C. I.: Projected Changes of Stratospheric Final Warmings in the Northern and Southern Hemispheres by CMIP5/6
875 Models, Clim Dyn, 56, 3353–3371, https://doi.org/10.1007/s00382-021-05647-6, 2021b.

Rao, J., Garfinkel, C. I., and White, I. P.: How Does the Quasi-Biennial Oscillation Affect the Boreal Winter Tropospheric Circulation in CMIP5/6 Models?, Journal of Climate, 33, 8975–8996, https://doi.org/10.1175/JCLI-D-20-0024.1, 2020a.

Rao, J., Garfinkel, C. I., White, I. P., and Schwartz, C.: The Southern Hemisphere Minor Sudden Stratospheric Warming in September 2019 and Its Predictions in S2S Models, Journal of Geophysical Research: Atmospheres, 125, e2020JD032 723,
880 https://doi.org/10.1029/2020JD032723, 2020b.

Richter, J. H., Solomon, A., and Bacmeister, J. T.: Effects of Vertical Resolution and Nonorographic Gravity Wave Drag on the Simulated Climate in the Community Atmosphere Model, Version 5, Journal of Advances in Modeling Earth Systems, 6, 357–383, https://doi.org/10.1002/2013MS000303, 2014.

Richter, J. H., Anstey, J. A., Butchart, N., Kawatani, Y., Meehl, G. A., Osprey, S., and Simpson, I. R.: Progress in Simu-
885 lating the Quasi-Biennial Oscillation in CMIP Models, Journal of Geophysical Research: Atmospheres, 125, e2019JD032 362, https://doi.org/10.1029/2019JD032362, 2020.

Richter, J. H., Glanville, A. A., Edwards, J., Kauffman, B., Davis, N. A., Jaye, A., Kim, H., Pedatella, N. M., Sun, L., Berner, J., Kim, W. M., Yeager, S. G., Danabasoglu, G., Caron, J. M., and Oleson, K. W.: A Subseasonal Earth System Prediction Framework with CESM2, submitted to Weather and Forecasting, 2021.

890 Saha, S., Nadiga, S., Thiaw, C., Wang, J., Wang, W., Zhang, Q., den Dool, H. M. V., Pan, H.-L., Moorthi, S., Behringer, D., Stokes, D., Peña, M., Lord, S., White, G., Ebisuzaki, W., Peng, P., and Xie, P.: The NCEP Climate Forecast System, Journal of Climate, 19, 3483–3517, https://doi.org/10.1175/JCLI3812.1, 2006.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., Chuang, H.-y., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., van den Dool, H., Zhang, Q., Wang, W., Chen, M., and Becker, E.: The NCEP Climate Forecast System Version 2,
895 Journal of Climate, 27, 2185–2208, https://doi.org/10.1175/JCLI-D-12-00823.1, 2014.

Schwartz, C., Garfinkel, C. I., Yadav, P., Chen, W., and Domeisen, D.: Stationary Waves and Upward Troposphere-Stratosphere Coupling in S2S Models, Weather and Climate Dynamics Discussions, pp. 1–25, https://doi.org/10.5194/wcd-2021-58, 2021.

Seviour, W. J. M., Mitchell, D. M., and Gray, L. J.: A Practical Method to Identify Displaced and Split Stratospheric Polar Vortex Events, Geophysical Research Letters, 40, 5268–5273, https://doi.org/10.1002/grl.50927, 2013.

Seviour, W. J. M., Hardiman, S. C., Gray, L. J., Butchart, N., MacLachlan, C., and Scaife, A. A.: Skillful Seasonal Prediction of the Southern Annular Mode and Antarctic Ozone, Journal of Climate, 27, 7462–7474, https://doi.org/10.1175/JCLI-D-14-00264.1, 2014.

Shaw, T. A. and Perlwitz, J.: The Impact of Stratospheric Model Configuration on Planetary-Scale Waves in Northern Hemisphere Winter, Journal of Climate, 23, 3369–3389, https://doi.org/10.1175/2010JCLI3438.1, 2010.

Shaw, T. A., Perlwitz, J., and Weiner, O.: Troposphere-Stratosphere Coupling: Links to North Atlantic Weather and Climate, Including Their Representation in CMIP5 Models, Journal of Geophysical Research: Atmospheres, 119, 5864–5880, https://doi.org/10.1002/2013JD021191, 2014.

Shepherd, T. G.: The Middle Atmosphere, Journal of Atmospheric and Solar-Terrestrial Physics, 62, 1587–1601, https://doi.org/10.1016/S1364-6826(00)00114-0, 2000.

Sigmond, M., Scinocca, J. F., Kharin, V. V., and Shepherd, T. G.: Enhanced Seasonal Forecast Skill Following Stratospheric Sudden Warmings, Nature Geosci, 6, 98–102, https://doi.org/10.1038/ngeo1698, 2013.

Son, S.-W., Lim, Y., Yoo, C., Hendon, H. H., and Kim, J.: Stratospheric Control of the Madden–Julian Oscillation, Journal of Climate, 30, 1909–1922, https://doi.org/10.1175/JCLI-D-16-0620.1, 2017.

Son, S.-W., Kim, H., Song, K., Kim, S.-W., Martineau, P., Hyun, Y.-K., and Kim, Y.: Extratropical Prediction Skill of the Subseasonal-to-Seasonal (S2S) Prediction Models, Journal of Geophysical Research: Atmospheres, 125, e2019JD031 273, https://doi.org/10.1029/2019JD031273, 2020.

Stockdale, T. N., Kim, Y.-H., Anstey, J. A., Palmeiro, F. M., Butchart, N., Scaife, A. A., Andrews, M., Bushell, A. C., Dobrynin, M., Garcia-Serrano, J., Hamilton, K., Kawatani, Y., Lott, F., McLandress, C., Naoe, H., Osprey, S., Pohlmann, H., Scinocca, J., Watanabe, S., Yoshida, K., and Yukimoto, S.: Prediction of the Quasi-Biennial Oscillation with a Multi-Model Ensemble of QBO-resolving Models, Quarterly Journal of the Royal Meteorological Society, n/a, https://doi.org/10.1002/qj.3919, 2020.

Taguchi, M.: Comparison of Subseasonal-to-Seasonal Model Forecasts for Major Stratospheric Sudden Warmings, Journal of Geophysical Research: Atmospheres, 123, 10 231–10 247, https://doi.org/10.1029/2018JD028755, 2018.

Thompson, D. W. J. and Wallace, J. M.: Annular Modes in the Extratropical Circulation. Part I: Month-to-Month Variability, Journal of Climate, 13, 1000–1016, https://doi.org/10.1175/1520-0442(2000)013<1000:AMITEC>2.0.CO;2, 2000.

Tripathi, O. P., Charlton-Perez, A., Sigmond, M., and Vitart, F.: Enhanced Long-Range Forecast Skill in Boreal Winter Following Stratospheric Strong Vortex Conditions, Environ. Res. Lett., 10, 104 007, https://doi.org/10.1088/1748-9326/10/10/104007, 2015.

Tyrrell, N. L. and Karpechko, A. Y.: Minimal Impact of Model Biases on Northern Hemisphere El Niño–Southern Oscillation Teleconnections, Weather and Climate Dynamics, 2, 913–925, https://doi.org/10.5194/wcd-2-913-2021, 2021.

Tyrrell, N. L., Karpechko, A. Y., and Rast, S.: Siberian Snow Forcing in a Dynamically Bias-Corrected Model, Journal of Climate, 33, 10 455–10 467, https://doi.org/10.1175/JCLI-D-19-0966.1, 2020.

Tyrrell, N. L., Koskentausta, J. M., and Karpechko, A. Y.: Sudden Stratospheric Warmings during El Niño and La Niña: Sensitivity to Model Biases, Weather and Climate Dynamics Discussions, pp. 1–20, https://doi.org/10.5194/wcd-2021-62, 2021.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H.-S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A. W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D.-J., Xiao, H., Zaripov, R., and Zhang, L.: The Subseasonal to Seasonal (S2S) Prediction Project Database, Bulletin of the American Meteorological Society, 98, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1, 2017.

Waugh, D. N. W.: Elliptical Diagnostics of Stratospheric Polar Vortices, Quarterly Journal of the Royal Meteorological Society, 123, 1725–1748, https://doi.org/10.1002/qj.49712354213, 1997.

940 White, I. P., Garfinkel, C. I., Gerber, E. P., Jucker, M., Aquila, V., and Oman, L. D.: The Downward Influence of Sudden Stratospheric Warmings: Association with Tropospheric Precursors, J. Climate, 32, 85–108, https://doi.org/10.1175/JCLI-D-18-0053.1, 2019.

White, I. P., Garfinkel, C. I., Gerber, E. P., Jucker, M., Hitchcock, P., and Rao, J.: The Generic Nature of the Tropospheric Response to Sudden Stratospheric Warmings, Journal of Climate, 33, 5589–5610, https://doi.org/10.1175/JCLI-D-19-0697.1, 2020.

Yoo, C. and Son, S.-W.: Modulation of the Boreal Wintertime Madden-Julian Oscillation by the Stratospheric Quasi-Biennial Oscillation,
945 Geophysical Research Letters, 43, 1392–1398, https://doi.org/10.1002/2016GL067762, 2016.

Zhao, M., Golaz, J.-C., Held, I. M., Guo, H., Balaji, V., Benson, R., Chen, J.-H., Chen, X., Donner, L. J., Dunne, J. P., Dunne, K., Durachta, J., Fan, S.-M., Freidenreich, S. M., Garner, S. T., Ginoux, P., Harris, L. M., Horowitz, L. W., Krasting, J. P., Langenhorst, A. R., Liang, Z., Lin, P., Lin, S.-J., Malyshev, S. L., Mason, E., Milly, P. C. D., Ming, Y., Naik, V., Paulot, F., Paynter, D., Phillipps, P., Radhakrishnan, A., Ramaswamy, V., Robinson, T., Schwarzkopf, D., Seman, C. J., Shevliakova, E., Shen, Z., Shin, H., Silvers,
950 L. G., Wilson, J. R., Winton, M., Wittenberg, A. T., Wyman, B., and Xiang, B.: The GFDL Global Atmosphere and Land Model AM4.0/LM4.0: 2. Model Description, Sensitivity Studies, and Tuning Strategies, Journal of Advances in Modeling Earth Systems, 10, 735–769, https://doi.org/10.1002/2017MS001209, 2018.